Nutri-bullets Hybrid: Consensual Multi-document Summarization

Citation: Shah, Darsh, Yu, Lili, Lei, Tao and Barzilay, Regina. 2021. "Nutri-bullets Hybrid: Consensual Multi-document Summarization." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

As Published: 10.18653/V1/2021.NAACL-MAIN.411

Publisher: Association for Computational Linguistics (ACL)

Persistent URL: https://hdl.handle.net/1721.1/143525

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Nutri-bullets Hybrid: Consensual Multi-document Summarization

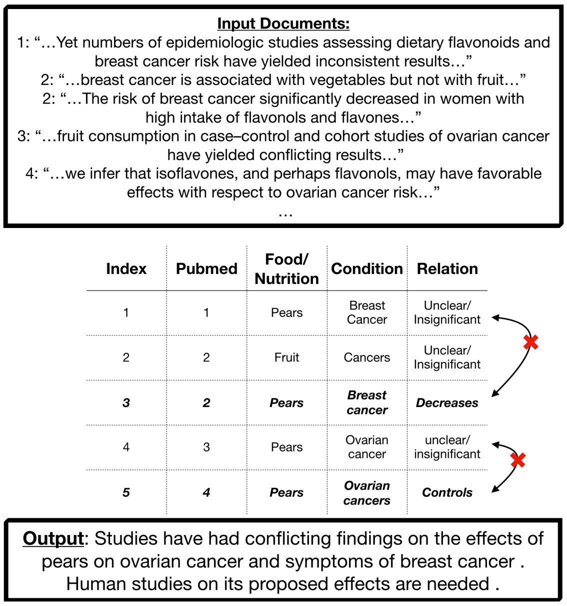
Darsh J Shah[1] **Lili Yu**[2] **Tao Lei**[2] **Regina Barzilay**[1]
[1]Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
[2]ASAPP, Inc.
darsh@csail.mit.edu liliyu@asapp.com tao@asapp.com regina@csail.mit.edu

Abstract

We present a method for generating comparative summaries that highlights similarities and contradictions in input documents. The key challenge in creating such summaries is the lack of large parallel training data required for training typical summarization systems. To this end, we introduce a hybrid generation approach inspired by traditional concept-to-text systems. To enable accurate comparison between different sources, the model first learns to extract pertinent relations from input documents. The content planning component uses deterministic operators to aggregate these relations after identifying a subset for inclusion into a summary. The surface realization component lexicalizes this information using a text-infilling language model. By separately modeling content selection and realization, we can effectively train them with limited annotations. We implemented and tested the model in the domain of nutrition and health – rife with inconsistencies. Compared to conventional methods, our framework leads to more faithful, relevant and aggregation-sensitive summarization – while being equally fluent.[1]

1 Introduction

Articles written about the same topic rarely exhibit full agreement. To present an unbiased overview of such material, a summary has to identify points of consensus and highlight contradictions. For instance, in the healthcare domain, where studies often exhibit wide divergence of findings, such comparative summaries are generated by human experts for the benefit of the general public.[2] Ideally, this capacity will be automated given a large number of relevant articles and continuous influx of new ones that require a summary update to keep it current. However, standard summarization architectures cannot be utilized for this task since the amount of comparative summaries is not sufficient for their training.



Input Documents:
1: "…Yet numbers of epidemiologic studies assessing dietary flavonoids and breast cancer risk have yielded inconsistent results…"
2: "…breast cancer is associated with vegetables but not with fruit…"
2: "…The risk of breast cancer significantly decreased in women with high intake of flavonols and flavones…"
3: "…fruit consumption in case–control and cohort studies of ovarian cancer have yielded conflicting results…"
4: "…we infer that isoflavones, and perhaps flavonols, may have favorable effects with respect to ovarian cancer risk…"
…

Index	Pubmed	Food/ Nutrition	Condition	Relation
1	1	Pears	Breast Cancer	Unclear/ Insignificant
2	2	Fruit	Cancers	Unclear/ Insignificant
3	***2***	***Pears***	***Breast cancer***	***Decreases***
4	3	Pears	Ovarian cancer	unclear/ insignificant
5	***4***	***Pears***	***Ovarian cancers***	***Controls***

Output: Studies have had conflicting findings on the effects of pears on ovarian cancer and symptoms of breast cancer . Human studies on its proposed effects are needed .

Figure 1: We consider the database extracted from four Pubmed studies on Pears and Cancer. The key facts (*bold*) and consensus (*contradiction*) are realized in the text generated by our model.

In this paper, we propose a novel approach to multi-document summarization based on a neural interpretation of traditional concept-to-text generation systems. Specifically, our work is inspired by the symbolic multi-document summarization system of (Radev and McKeown, 1998) which produces summaries that explicitly highlight agreements, contradictions and other relations across input documents. While their system was based on human-crafted templates and thus limited to a narrow domain, our approach learns different components of the generation pipeline from data.

To fully control generated content, we frame the task of comparative summarization as concept-to-text generation. As a pre-processing step, we ex-

[1] Our code and data is available at https://github.com/darsh10/Nutribullets

[2] Examples include https://www.healthline.com and https://foodforbreastcancer.com.

Proceedings of the 2021 Conference of the North American Chapter of the
Association for Computational Linguistics: Human Language Technologies, pages 5213–5222
June 6–11, 2021. ©2021 Association for Computational Linguistics

tract pertinent entity pairs and relations (see Figure 1) from input documents. The *Content Selection* component identifies the key tuples to be presented in the final output and establishes their comparative relations (e.g., consensus) via aggregation operators. Finally, the *surface realization* component utilizes a text-infilling language model to translate these relations into a summary. Figure 1 exemplifies this pipeline, showing selected key pairs (marked in bold), their comparative relation – *Contradiction* (rows 1 &3 and rows 4&5 conflict), and the final summary.[3]

This generation architecture supports refined control over the summary content, but at the same time does not require large amounts of parallel data for training. The latter is achieved by separately training content selection and content realization components. Since the content selection component operates over relational tuples, it can be robustly trained to identify salient relations utilizing limited parallel data. Aggregation operators are implemented using simple deterministic rules over the database where comparative relations between different rows are apparent. On the other hand, to achieve a fluent summary we have to train a language model on large amounts of data, but such data is readily available.

In addition to training benefits, this hybrid architecture enables human writers to explicitly guide content selection. This can be achieved by defining new aggregation operators and including new inference rules into the content selection component. Moreover, this architecture can flexibly support other summarization tasks, such as generation of updates when new information on the topic becomes available.

We apply our method for generating summaries of Pubmed publications on nutrition and health. Typically, a single topic in this domain is covered by multiple studies which often vary in their findings making it particularly appropriate for our model. We perform extensive automatic and human evaluation to compare our method against state-of-the-art summarization and text generation techniques. While seq2seq models receive competent fluency scores, our method performs stronger on task-specific metrics including *relevance*, *content faithfulness* and *aggregation cognisance*. Our method is able to produce summaries that receive an absolute 20% more on aggregation cognisance, an absolute 7% more on content relevance and 7% on faithfulness to input documents than the next best baseline in traditional and update settings.

2 Related Work

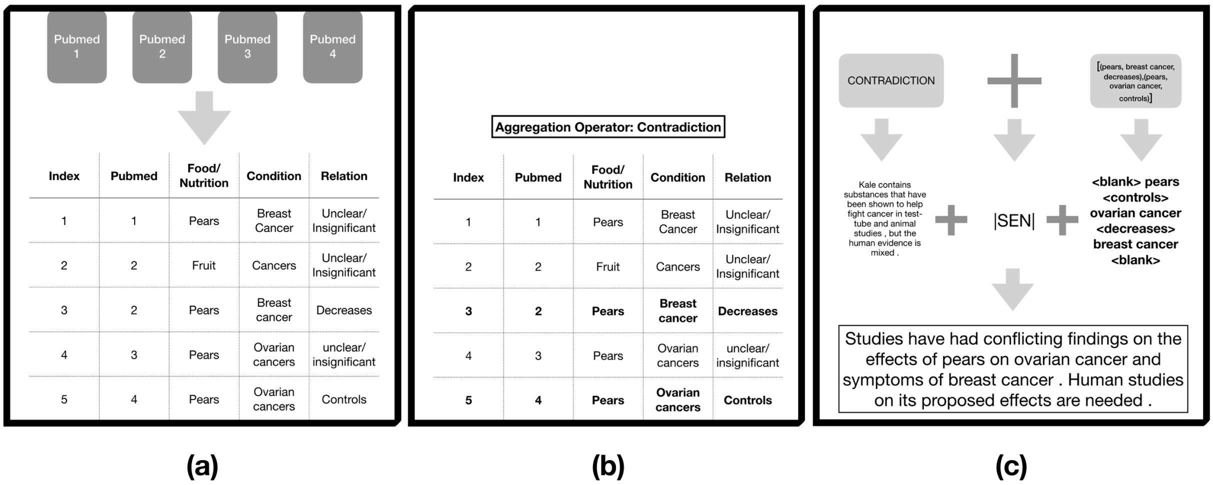
Text-to-text Summarization Neural sequence-to-sequence models (Rush et al., 2015; Cheng and Lapata, 2016; See et al., 2017) for document summarization have shown promise and have been adapted successfully for multi-document summarization (Zhang et al., 2018; Lebanoff et al., 2018; Baumel et al., 2018; Amplayo and Lapata, 2019; Fabbri et al., 2019). Despite producing fluent text, these techniques may generate false information which is not faithful to the original inputs (Puduppully et al., 2019; Kryściński et al., 2019), especially in low resource scenarios. In this work, we are interested in producing faithful and fluent text cognizant of aggregation amongst input documents, where few parallel examples are available.

Recent language modeling approaches (Devlin et al., 2018; Stern et al., 2019; Shen et al., 2020; Donahue et al., 2020) can also be extended for text completion. Our work is a text-infilling language model where we generate words in place of relation specific blanks to produce a faithful summary.

Prior work (Mueller et al., 2017; Fan et al., 2017; Guu et al., 2018) on text generation also control aspects of the produced text, such as style and length. While these typically utilize tokens to control the modification, using prototypes to generate text is also very common (Guu et al., 2017; Li, 2018; Shah et al., 2019). In this work, we utilize aggregation specific prototypes to guide aggregation cognizant surface realization.

Data-to-text Summrization Traditional approaches for data-to-text generation have operated on symbolic data from databases. McKeown and Radev (1995); Radev and McKeown (1998); Barzilay et al. (1998) introduce two components of content selection and surface realization. Content selection identifies and aggregates key symbolic data from the database which can then be realized into text using templates. Unlike modern data-to-text systems (Wiseman et al., 2018; Puduppully et al., 2019; Sharma et al., 2019; Wenbo et al., 2019) these approaches capture document consensus and aggregation cognisance. While the neural approaches alleviate the need for human intervention, they do need an abundance of parallel data,

[3] We compare the selected content with other entries in the database, identifying two contradictions.


Pubmed 1
Pubmed 2
Pubmed 3
Pubmed 4
(a)
(b)
(c)
CONTRADICTION
[(pears, breast cancer, decreases),(pears, ovarian cancer, controls)]
Kale contains substances that have been shown to help fight cancer in test-tube and animal studies , but the human evidence is mixed .
|SEN|
<blank> pears <controls> ovarian cancer <decreases> breast cancer <blank>

Index	Pubmed	Food/ Nutrition	Condition	Relation
1	1	Pears	Breast Cancer	Unclear/ Insignificant
2	2	Fruit	Cancers	Unclear/ Insignificant
3	2	Pears	Breast cancer	Decreases
4	3	Pears	Ovarian cancers	unclear/ insignificant
5	4	Pears	Ovarian cancers	Controls

Aggregation Operator: Contradiction

Index	Pubmed	Food/ Nutrition	Condition	Relation
1	1	Pears	Breast Cancer	Unclear/ Insignificant
2	2	Fruit	Cancers	Unclear/ Insignificant
3	**2**	**Pears**	**Breast cancer**	**Decreases**
4	3	Pears	Ovarian cancers	unclear/ insignificant
5	**4**	**Pears**	**Ovarian cancers**	**Controls**

Studies have had conflicting findings on the effects of pears on ovarian cancer and symptoms of breast cancer . Human studies on its proposed effects are needed .

Figure 2: Illustrating the flow of our *Nutribullets Hybrid* system. In this example, our model takes in four Pubmed studies to produce a database (a). The *Content Selection* model selects two tuples (bold) and identifies the aggregation operator as Contradiction (b). Finally, the *Surface Realization* model takes in the tuples and aggregation operator to produces a summary which is faithful to input entities and aggregation cognizant (c).

which are typically from one source only. Hence, modern techniques do not deal with input documents' consensus in low resource settings.

3 Method

Our goal is to generate a text summary y for a food from a pool of multiple scientific abstracts X. In this section, we describe the framework of our *Nutribullets Hybrid* system, illustrated in Figure 2.

3.1 Overview

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 2).

Notations: For N input documents, we collect $X_{\mathcal{G}} = \{\mathcal{G}_p^x\}_{p=1}^N$, a database of entity-entity relations $\mathcal{G}_p^x$. $\mathcal{G}_p = (e_1^k, e_2^k, r^k)_{k=1}^K$ is a set of K tuples of two entities e_1, e_2 and their relation r. r represents relations such as the effect of a nutrition entity e_1 on a condition e_2 (see Table 2).[4]We have raw text converted into symbolic data.

Similarly, we denote the corpus of summaries as $Y = \{(y_m, \mathcal{G}_m^y, O_m^y)_{m=1}^M\}$, where y_m is a concise summary, $\mathcal{G}_m^y$ is the set of entity-entity relation tuples and O_m^y is the realized aggregation, in M data points.

Modeling: Joint learning of content selection, information aggregation and text generation for multi-document summarization can be challenging. This is further exacerbated in our technical domain with few parallel examples and varied consensus amongst input documents. To this end, we propose a solution using Content Selection and Aggregation and Surface Realization models.

Raw text from N input documents is converted into a mini-database $X_{\mathcal{G}}$ of relation tuples. The content selection and aggregation model operates on such symbolic data. We use $X_{\mathcal{G}}$ and Y to train the content selection model. During inference, we identify from $X_{\mathcal{G}}$ a subset C of content to present in the final output. In order to produce a summary cognizant of consensus amongst inputs, we identify the aggregation operator O based on C and other relevant tuples in $X_{\mathcal{G}}$.

The surface realization model produces a relevant, faithful and aggregation cognizant output. The model is trained only using Y. During inference, the model realizes text using the selected content C and the aggregation operator O.

3.2 Content Selection and Aggregation

Our content selection model takes a mini-database of entity-entity relation tuples $X_{\mathcal{G}}$ as input, and outputs the key tuples C and the aggregation operator O.

Content selection and aggregation consists of two parts – (i) identifying key content $P(C|X_{\mathcal{G}})$ and (ii) subsequently identifying the aggregation operator O using $C, X_{\mathcal{G}}$.

Content Selection Identifying key content in-

[4] We train an entity tagger and relation classifier to predict $\mathcal{G}$ and also for computing knowledge based evaluation scores. More details on models and results are shared later.

Aggregation Operator	Deterministic Rule
Under-Reported	|Pubmed Studies| < Threshold
Population Scoping	|Specific Population| < Threshold
Contradiction	$(e_1^m == e_1^n \&\& e_2^m == e_2^n \&\& r^m ! = r^n)$ for any two tuples m, n from different studies
Agreement	None of the Above

Table 1: Deterministic Rules to identify the Aggregation Operator.

volves selecting important, diverse and representative tuples from a database. While clustering and selecting from the database tuples is a possible solution, we model our content selection as a finite Markov decision process (MDP). This allows for an exploration of different tuple combinations while incorporating delayed feedback from various critical sources of supervision (similarity with target tuples, diversity amongst selected tuples etc). We consider a multi-objective reinforcement learning algorithm (Williams, 1992) to train the model. Our rewards (Eq. 2) allow for the selection of informative and diverse relation tuples.

The MDP's state is represented as $s_t = (t, \{c_1, \ldots, c_t\}, \{z_1, z_2, ..., z_{m-t}\})$ where t is the current step, $\{c_1, \ldots, c_t\}$ is the content selected so far and $\{z_1, z_2, ..., z_{m-t}\}$ is the remaining entity-entity relation tuples in the m-sized database. The action space is all the remaining tuples plus one special token, $Z \cup \{\mathit{STOP}\}$.[5] The number of actions is equal to $|m - t| + 1$. As the number of actions is variable yet finite, we parameterize the policy $\pi_\theta(a|s_t)$ with a model f which maps each action and state (a, s_t) to a score, in turn allowing a probability distribution over all possible actions using softmax. At each step, the probability that the policy selects z_i as a candidate is:

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

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, $cos(u, v) = \frac{u \cdot v}{||u|| \cdot ||v||}$ is the cosine similarity of two vectors and f is a feed-forward neural network with non-linear activation functions that outputs a scalar score for each action a.

The selection process starts with Z. Our module iteratively samples actions from $\pi_\theta(a|s_t)$ until selecting *STOP*, ending with selected content C and a corresponding reward. We can even allow for the selection of partitioned tuple sets by adding an extra action of "NEW LIST", which allows the model to include subsequent tuples in a new group.

We consider the following individual rewards:

- $\mathcal{R}_e = \sum_{c \in C} cos(\hat{e_{1c}}, \hat{e_{1y}}) + cos(\hat{e_{2c}}, \hat{e_{2y}})$ is the cosine similarity of the structures of the selected content C with the structures present in the summary y (each 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 tuples.
- r_p is a small penalty for each action step to encourage concise selection.

The multi-objective reward is computed as

$$\mathcal{R} = w_e \mathcal{R}_e + w_d \mathcal{R}_d - |C| r_p, \quad (2)$$

where w_e, w_d and r_p are hyper-parameters.

During training the model is updated based on the rewards. During inference the model selects an ordered set of key and diverse relation tuples corresponding to appropriate health conditions.

Consensus Aggregation Identifying the consensus amongst the input documents is critical in our multi-document summarization task. We model the aggregation operator of our *Content Selection* using simple one line deterministic rules as shown in Table 1. The rules are applied to the key C entity-entity relation pairs in context of X_G. In our example in Figure 1, O is Contradiction because of rows 1&3 and rows 4&5 (rows 1&3 only would also make it Contradiction).

3.3 Surface Realization

The surface realization model $P(y|O, C)$, performs the critical task of generating a summary guided by both the entity-entity relation tuples C and the aggregation operator O. The model allows for robust, diverse and faithful summarization compared to traditional template and modern seq2seq approaches.

[5] STOP and NEW LIST get special embeddings.

Relation Type	e_1	e_2	r	**Example**
Causing	Food, Nutrition	Condition	Increase, Decrease, Satisfy, Control, Unclear/Insignificant	(tart cherry juice, melatonin levels, increase), (water, daily fluid needs, satisfy)
Containing	Food, Nutrition	Nutrition	Contain	(blueberries, antioxidants, contain)

Table 2: Details of entity-entity relationships that we study and some examples of (e_1, e_2, r)

We propose to model this process as a prototype-driven text infilling task. The entities from C are used as fixed tokens with relations as special blanks in between these entities. This is prefixed by a prototype summary corresponding to O. For the example shown in Figure 2, we concatenate using $|SEN|$ a randomly sampled contradictory summary *"Kale contains substances ... help fight cancer ... but the human evidence is mixed ."* to C *"<blank> pears <controls> ovarian cancer <decreases> breast cancer <blank>"*. The infilling language model produces text corresponding to relations between entities while maintaining an overall structure which is cognizant of O. [6]

The model is trained on the few sample summaries from the training set using $\mathcal{G}_m^y$ and O_m^y to produce y_m. Providing aggregation and content guidance during generation alleviates the low-resource issue.

4 Summary and Update Setting

In this section we describe the setting of summary updates. In a real world setting, we would often receive new input documents such as scientific studies about the same subject which necessitate a change in an old summary.

In context of our food and health summarization task, the goal is to update an old summary about a food and health condition on receiving results from new scientific studies from Pubmed. Our model can accommodate this scenario fairly easily. We describe the minor changes to the *Content Selection and Aggregation* and *Surface Realization* models for such a setting.

We are provided an original summary and can extract it's content C' and can also construct the mini-database X_G from the text of the new documents. We identify the aggregation between the new studies' X_G and original summary's content C' first. Depending on the aggregation identified, corresponding content C is selected from X_G. For instance, in case of a contradiction, we are keen on identifying content leading to this contradiction. The subsequent *Surface Realization* is dependent on O, the selected C and the C' present in the original summary ($P(y|O, C + C')$).

5 Experiments

Dataset We utilize a real world dataset for Food and Health summaries, crawled from `https://www.healthline.com/nutrition` (Shah et al., 2021). The HealthLine dataset consists of scientific abstracts as inputs and human written summaries as outputs. The dataset consists of 6640 scientific abstracts from Pubmed, each averaging 327 words. The studies in these abstracts are cited by domain experts when writing summaries in the Healthline dataset, forming natural pairings of parallel data. Individual summaries average 24.5 words and are created using an average of 3 Pubmed abstracts. Each food has multiple bullet summaries, where each bullet typically talks about a different health impact (hydration, diabetes etc). We assign each food article randomly into one of the train, development or test splits. Entity tagging and relation classification annotations are provided for the Pubmed abstracts and the healthline summaries.

Settings: We consider three settings.

1. Single Issue: We use the individual food and health issue summaries as a unique instance of food and single issue setting. We split 1894 instances 80%,10%,10% to train, dev and test.

2. Multiple Issues: We group each food's article Pubmed abstract inputs and multiple summary outputs as a single parallel instance. 464 instances are split 80%,10%,10% to train, dev and test.

3. Summary Update: We consider two kinds of updates – new information is fused to an existing summary and new information contradicts an existing summary. For fusion we consider single issue summaries that have multiple conditions from different Pubmed studies (bananas + low blood pressure from one study and bananas + heart health from another study). We partition the Pubmed

[6] Summaries in our training data are labelled with O_m^y as belonging to one of the four categories of *Under-reported, Population Scoping, Contradiction or Agreement* to accommodate such training.

	Automatic Evaluation				Human Scores	
MODEL	ROUGEL	KG(G)	KG(I)	AG	RELEVANCE	FLUENCY
Copy-gen	0.12	0.21	0.50	0.64	1.93	1.89
GraphWriter	0.14	0.03	0.69	0.64	1.86	2.76
Entity Data2text	0.16	0.13	0.57	0.67	2.03	3.43
Transformer	**0.20**	0.21	0.64	0.67	2.66	**3.76**
Ours	0.18	**0.30**	**0.76**	**0.89**	**3.03**	3.46

Table 3: Automatic evaluation – Rouge-L score (RougeL), KG in gold(G), KG in input(I) and Aggregation Cognisance (Ag) in our model and various baselines in the single issue setting, is reported. Human evaluation on Relevance and Fluency, on 1-4 Likert scale from 3 annotators, is also reported. The best results are in **bold**.

studies to stimulate an update. The contradictory update setting is where we artificially introduce conflicting results in the input document set so that the aggregation changes from Agreement to Contradictory. We have a total of 103 test instances. All models are trained atop of Single issue data.

Evaluation We evaluate our systems using the following automatic metrics. *Rouge* is an automatic metric used to compare the model output with the gold reference (Lin, 2004). *KG(G)* computes the number of entity-entity pairs with a relation in the gold reference, that are generated in the output.[7] 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. *Aggregation Cognisance (Ag)* measures the accuracy of the model in producing outputs which are cognizant of the right aggregation from the input, (Under-reported, Contradiction or Agreement). We use a rule-based classifier to identify the aggregation implied by the model output and compare it to the actual aggregation operator based on the input Pubmed studies.

In addition to automatic evaluation, we have human annotators score our models on relevance and fluency. Given a reference summary, *relevance* indicates if the generated text shares similar information. *Fluency* represents if the generated text is grammatically correct and written in well-formed English. Annotators rate relevance and fluency on a 1-4 likert scale (Albaum, 1997). We have 3 annotators score every data point and report the average across the scores.

Baselines In order to demonstrate the effectiveness of our method, we compare it against text2text and data2text state-of-the-art (*sota*) methods.

Copy-gen (Text2text): See et al. (2017) is a *sota* technique for summarization, which can copy from the input or generate words.

Transformer (Text2text): Hoang et al. (2019) is a summarization system using a pretrained Transformer.

GraphWriter (Data2text): Koncel-Kedziorski et al. (2019) is a graph transformer based model, which generates text using a seed title and a knowledge graph. Takes the database X_G as input.

Entity (Data2text): Puduppully et al. (2019) is an entity based data2text model, takes X_G as input.

Implementation Details Our policy network is a three layer feedforward neural network. We use a Transformer (Vaswani et al., 2017) implementation for Surface Realization. We train an off-the-shelf Neural CRF tagger (Yang and Zhang, 2018) for entity extraction. We use BERT (Devlin et al., 2018) based classifiers to predict the relation between two entities in a text trained using crowdsourced annotations from (Shah et al., 2021). Futher implementation details can be found in A.

6 Results

In this section, we describe the performance of our *Nutribullet Hybrid* system and baselines on summarization and summary updates. We report empirical results , human evaluation and present sample outputs, highlighting the benefits of our method.

Single and Multi-issues Summarization: We describe the results on the task of generating summaries. Table 3 presents the automatic evaluation results for the food and single issue summarization task. High KG(I) and KG(G) scores for our method indicate that the generated text is faithful to input entities and relevant. In particular, a high Aggregation Cognisance (Ag) score indicates that our model generates summaries which are cog-

[7] 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). A cosine score exceeds a threshold of 0.7 is set (minimize false positives) to identify a match.

Transformer (baseline)
* Whole - grain cereals may protect against obesity , diabetes and certain cancers. However , more research is needed .
* Whole grains , such as mozambican grass , are safe to eat with no serious side effects .
* Whole - grain cereals may protect against obesity , diabetes and certain cancers. However , more research is needed .
* Whole grains , such as blueberries , are likely safe to eat with no serious side effects .
* Whole grains are safe to eat. However , people with type 2 diabetes should avoid whole grains .
* Whole grains are lower in carbs than whole grains , making them a good choice for people with type 2 diabetes.
Our Method
* Whole grains has been shown to lower weight gain and improve various type 2 diabetes risk factors .
* Whole grains has been shown to lower insulin resistance and improve various cancer risk factors .
* Whole grains has been linked to several other potential health benefits , such as improved CVD risk , eyesight , and memory. However , more studies are needed to draw stronger conclusions.
* There is some evidence , in both animals and humans , that whole grains can reduce mortality by regulating the hormone ghrelin.

Table 4: Example outputs of our model and the Transformer baseline for a multi-issues summary. Trained on limited parallel data, the Transformer baseline produces repetitive text with factual inaccuracies, while our method is able to provide more accurate and diverse summarization.

Model	KG(G)	KG(I)
Copy-gen	0.43	0.69
Transformer	0.33	0.73
Ours	**0.5**	**0.90**

Table 5: KG in gold(G) and KG in input(I) in our model and baselines in the food and multi-issues setting . The best results are in **bold**.

nizant of the varying degrees of consensus in the input Pubmed documents. Compared to other baselines we also receive a competitive score on the automatic Rouge metric, beating Copy-gen, Entity Data2text and GraphWriter baselines while falling short (by 1.7%) of the Transformer baseline. The baselines, especially Transformer, tend to produce similar outputs for different inputs (see Table 4). Since a lot of these patterns are learned from the human summaries, Transformer receives a high Rouge score. However, as in the low resource regime, the baseline does not completely capture the content and aggregation, it fails to get a very high KG(G) or Ag score. A similar trend is observed for the other baselines too, which in this low resource regime produce a lot of false information, reflected in their low KG(I) scores.

Human evaluation, conducted by considering scores,on a 1-4 Likert scale, from three annotators for each instance, shows the same pattern. Our model is able to capture the most relevant information, when compared against the gold summaries while producing fluent summaries. The Transformer baseline produces fluent summaries, which are not as relevant. The performance is poorer for the Copy-gen, Entity Data2text and GraphWriter models.

In the multi-issues setting, the baselines access the gold annotations with respect to the input documents' clustering. Our model conducts the extra task of grouping the selected tuples, using the "New List" action. Our model performs better than the baselines on both the KG(I) and KG(G) metrics as seen in Table 5. Again, the pattern of producing very similar and repetitive sentences hurts the baselines. They fail to cover different issues and tend to produce false information, in this low resource setting. Our model scores an 7% higher on KG(G) and 17% higher on KG(I) compared to the next best performance, in absolute terms. Table 4 shows the comparison between the outputs produced by our method and the Transformer baseline on the benefits of whole-grains. Our method conveys more relevant, factual and organized information in a concise manner.

	Fusion Update	Contradictory Update
Model	KG(G)	Ag
Copy-gen	0.16	0.50
GraphWriter	0.0	0.50
Entity Data2text	0.16	0.50
Transformer	0.16	0.46
Ours	**0.33**	**0.76**

Table 6: The middle column shows KG in gold(G) in our model and baselines for fusion updates . The last column shows Aggregation Cognisance (Ag) in our model and baselines in the contradictory update setting. The best results are in **bold**.

Summary Update: We study the efficacy of our model to fuse information in existing summaries on receiving new Pubmed studies. As the KG(G) metric in 6 shows, our model is able to select and fuse more relevant information. Table 7 shows two examples of summaries on flaxseeds where our model successfully fuses new information.

Table 6's last column presents the automatic

Old Summary	Flax seeds contain a group of nutrients called lignans , which have powerful antioxidant and estrogen properties .
New Inputs	(i):"...current overall evidence indicates that FS and its components are effective in the risk reduction and treatment of breast cancer and safe for consumption by breast cancer patients..." (ii): "...Consumption of flaxseed was associated with a significant reduction in breast cancer risk as was consumption of flax bread ..." (iii): "...a flaxseed-supplemented, fat-restricted diet may affect the biology of the prostate and associated biomarkers..."
Copy-gen	Avocados may help fight cancer risk, boost inflammation. In a pasteurized called polyphenols, which may aid weight loss.
Transformer	Flaxseed oil is high in antioxidants that may help reduce the risk of several chronic diseases .
Ours	Flax seeds are rich in antioxidant , especially through lignans. They contain beneficial nutrients which can help protect your body against certain types of breast cancer .
Old Summary	Flax seeds, high in fiber, can be a beneficial addition to the diet of people with diabetes .
New Input	"...showed fasting blood sugar in the experimental group decreased...the total cholesterol reduced...Results showed a decrease in low-density lipoprotein cholesterol...The study demonstrated the efficacy of flax gum in the blood biochemistry profiles of type 2 diabetes."
Copy-gen	Eating apart has been linked to increased growth cholesterol, and cholesterol levels. However, more studies are needed to confirm possible effect.
Transformer	Flaxseed extract may help lower blood sugar levels .
Ours	Flax seeds are high in fiber , which is beneficial for people with diabetes and associated with a reduced low-density lipoprotein cholesterol .

Table 7: Example outputs of our model and baselines for a summary update upon receiving new information about flaxseeds + cancer and flaxseeds + cholesterol, respectively. Our model maintains old information and updates accurately. In the cholesterol case, Transformer adds new information but misses the old information.

evaluation results to demonstrate the efficacy of maintaining Aggregation Cognisance (Ag), which is critical when updating summaries on receiving contradictory results. The high performance in this update setting demonstrates the *Surface Realization* model's ability to produce aggregation cognizant outputs, in contrast to the baselines that do not learn this reasoning in a low resource regime.

Analysis: Information Extraction and Content Aggregation Information extraction is the critical first step performed for the input documents in order to get symbolic data for content selection and aggregation. To this end, we report the performance of the information extraction system, which is composed of two models – entity extraction and relation classification. As reported in Table 8, the entity extraction model, a crf-based sequence tagging model, receives a token-level F1 score of 79%. The relation classification model, a BERT based text classifier, receives an accuracy of 69%.

The performance of the information extraction models is particularly important for the content aggregation sub-task. In order to analyse this quantitatively, we perform manual analysis of the 179 instances in the dev set and compare them to the system identified aggregation – information extraction followed by the deterministic rules in Table 1. Given the simplicity of our rules, system's 78% accuracy in Table 8 is acceptable. Deeper analysis shows that the performance is lowest for Population Scoping and Contradiction with an accuracy of 52% and 56% respectively. The performance of Population Scoping being low is down predominantly to the simplicity of the rules. Most mistakes occur when the input studies are review studies that don't mention any population but analyze results from several past work. Contradiction suffers because of the information extraction system and stronger models for the same should be able to alleviate the errors.

Task	Performance
Entity Extraction	0.79
Relation Classification	0.69
Aggregation Operator Identification	0.78

Table 8: Performance of our information extraction system and its impact on content aggregation.

7 Conclusion

While modern models produce fluent text in multi-document summarization, they struggle to capture the consensus amongst the input documents. This inadequacy – magnified in low resource domains, is addressed by our model. Our model is able to generate robust summaries which are faithful to content and cognizant of the varying consensus in the input documents. Our approach is applicable in summarization and textual updates. Extensive experiments, automatic and human evaluation underline its impact over state-of-the-art baselines.

Acknowledgements

We thank the MIT NLP group and the reviewers for their helpful discussion and comments. This work is supported by DSO grant DSOCO1905.

References

Gerald Albaum. 1997. The likert scale revisited. *Market Research Society. Journal.*, 39(2):1–21.

Reinald Kim Amplayo and Mirella Lapata. 2019. Informative and controllable opinion summarization. *arXiv preprint arXiv:1909.02322*.

Regina Barzilay, Daryl McCullough, Owen Rambow, Jonathan DeCristofaro, Tanya Korelsky, and Benoit Lavoie. 1998. A new approach to expert system explanations. In *Natural Language Generation*.

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

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. *arXiv preprint arXiv:2005.05339*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Kelvin Guu, Panupong Pasupat, Evan Liu, and Percy Liang. 2017. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1051–1062, Vancouver, Canada. Association for Computational Linguistics.

Andrew Hoang, Antoine Bosselut, Asli Celikyilmaz, and Yejin Choi. 2019. Efficient adaptation of pretrained transformers for abstractive summarization. *arXiv preprint arXiv:1906.00138*.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv*, pages arXiv–1910.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. *arXiv preprint arXiv:1808.06218*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yanpeng Li. 2018. Learning features from cooccurrences: A theoretical analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2846–2854, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. *arXiv preprint arXiv:1906.03221*.

Dragomir R. Radev and Kathleen R. McKeown. 1998. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Darsh J Shah, Tal Schuster, and Regina Barzilay. 2019. Automatic fact-guided sentence modification. *arXiv preprint arXiv:1909.13838*.

Darsh J Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021. Nutri-bullets: Summarizing health studies by composing segments. *arXiv preprint arXiv:2103.11921*.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. *arXiv preprint arXiv:2002.03079*.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. *arXiv preprint arXiv:1902.03249*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang Wenbo, Gao Yang, Huang Heyan, and Zhou Yuxiang. 2019. Concept pointer network for abstractive summarization. *arXiv preprint arXiv:1910.08486*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Jie Yang and Yue Zhang. 2018. Ncrf++: An open-source neural sequence labeling toolkit. *arXiv preprint arXiv:1806.05626*.

Jianmin Zhang, Jiwei Tan, and Xiaojun Wan. 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*, pages 381–390, Tilburg University, The Netherlands. Association for Computational Linguistics.