SUMPUBMED: Summarization Dataset of PubMed Scientific Articles

Anonymous ACL-IJCNLP submission

Abstract

Most earlier work on text summarization is carried out on news article datasets. The summary in these datasets is naturally located at the beginning of the text. Hence, a model can spuriously utilize this correlation for summary generation instead of truly learning to summarize. To address this issue, we constructed a new dataset, SUMPUBMED, using scientific articles from the PubMed archive. We conducted a human analysis of summary coverage, redundancy, readability, coherence, and informativeness on SUMPUBMED. SUMPUBMED is challenging because (a) the summary is distributed throughout the text (not-localized on top), and (b) it contains rare domain-specific scientific terms. We observe that seq2seq models that adequately summarize news articles struggle to summarize SUMPUBMED. Thus, SUMPUBMED opens new avenues for the future improvement of models as well as the development of new evaluation metrics.

1 Introduction

Most of the existing summarization datasets, i.e., CNN Daily Mail and DUC are news article datasets. That is, the article acts as a document, and the summary is a short (10-15 lines) manually written highlight (i.e., headlines). In many cases, these highlights have significant lexical overlap with the few lines at the top of the article. Thus, any model which can extract the top few lines, e.g., extractive methods, performs adequately on these datasets.

However, the task of summarization is not merely limited to short-length news articles. One could also summarize long and complex documents such as essays, research papers, and books. In such cases, an extractive approach will most likely fail. For successful summarization on these documents, one needs to (a) find information from the distributed (non-localized) locale in the large text, (b) perform paraphrasing, simplifying, and shortening of longer sentences and (c) combine information from multiple sentences to generate the summary. Hence, an abstractive approach will perform better on such large documents.

One obvious source that contains such complex documents is the MEDLINE biomedical scientific articles, which are publicly available. Furthermore, these articles are accompanied by abstracts and conclusions which summarize the documents. Therefore, we constructed a scientific summarization dataset from pre-processed PubMed articles, named SUMPUBMED. In comparison to the previous news-article based datasets, SUMPUBMED documents are longer, and the corresponding summaries cannot be extracted by selecting a few sentences from fixed locations in the document.

The dataset, along with associated scripts, is available at anonymous_for_submission. Our contributions in this paper are:

- We created a new scientific summarization dataset, SUMPUBMED, which has longer text documents and summaries with non-localized information from documents.
- We analyzed the quality of summaries in SUMPUBMED on the basis of four parameters: readability, coherence, non-repetition, and informativeness using human evaluation.
- We evaluated several extractive, abstractive (seq2seq), and hybrid summarization models on SUMPUBMED. The results show that SUMPUBMED is more challenging compared to the earlier news-based datasets.
- Lastly, we showed that the standard summarization evaluation metric, ROUGE (Lin, 2004), correlates poorly with human evaluations on SUMPUBMED. This indicates the

need for a new evaluation metric for the scientific summarization task.

2 SumPubMed Creation

SumPubMed is created from PubMed biomedical research papers, which has 26 million documents. The documents are sourced from diverse literature, including MEDLINE, life science journals, and online books. For SumPubMed creation we took $33,772$ documents from Bio Med Central (BMC). BMC incorporates research papers related to medicine, pharmacy, nursing, dentistry, health care, health services, etc.

The research documents in BMC contain two subsections: *Front* and *Body*. The front part of the document is basically the abstract and taken as the gold summary. The body part which is taken as the main document contains three subsections: background, results, and conclusion. The average word count in a PubMed article is $4,200$ words distributed within 300 lines. To attain a smaller manageable document size, we performed extensive preprocessing to reduce text. During preprocessing, the non-textual content from the text was removed by: (a) replacing citations and digits in the content with <*cit*> and <*dig*> labels, (b) removing figures, tables, signatures, subscripts, superscripts, and their associated text (e.g., captions), and (c) removing the acknowledgments and references from the text. All the preprocessing was done on a sentence level utilizing the Python regex library.[1] After preprocessing, we convert the final document to an *XML* format and use the *SAX* parser to parse it. The difference between *SAX* and *DOM* parsers is noted in Appendix B.[2] An example of the front part, body part, and the XML file formed from the pre-processed text is shown in Appendix E.

Versions of SumPubMed: We maintained three versions of SumPubMed with varying degrees of preprocessing, a) XML, b) Raw Text, and c) Noun-phrases. Details of each version are as follows: (a) In the *XML* version, we exported the whole dataset into a single *XML* file, (b) The Raw Text version is obtained after preprocessing when removing non-textual context is completed, followed by *XML* parsing. (c) In the Noun phrases version, we processed the raw text version further to ensure that the summary and the text have the same named entities. We found that standard Name Entity Recognition (*NER*) (Finkel et al., 2005) and Biomedical Named Entity Recognizer (*ABNER*) (Settles, 2005) fail to pick the scientific named entities correctly.[3] Therefore, we use a simple heuristic of noun intersection between summary and main-text noun phrases to obtain plausible entity sets. This produced a shorter version of both the text and the summary than the original pair.

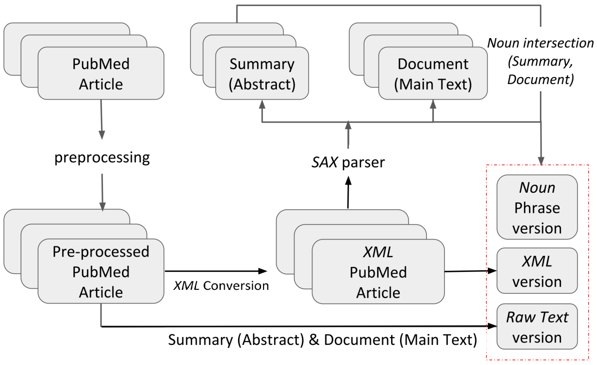

Figure 1: SumPubMed creation pipeline.

Version	Avg. Stats	Summary	Article
Raw Text version	Words	277	4227
	Sents	14	203
Noun Phrase version	Words	223	1578
	Sents	10	57
Hybrid version	Words	223	1891
	Sents	10	71

Table 1: Average number of sentences and words in the abstract and text in the three SumPubMed versions

The SumPubMed versions statistics is given in Table 1. The SumPubMed overall creation pipeline is shown in Figure 1.

3 Human Annotation of SumPubMed

Inspired from work on human evaluation of summaries by Friedrich et al. (2014), we distributed 50 randomly chosen summaries from the noun-phrase versions of SumPubMed to 10 expert annotators (graduate students) such that we have 3 annotation for each summary. We asked these human-annotators to rate the summaries on a scale of 1 to 10. We created different document files, each having 10 pairs of summaries where we randomly shuffled between reference and generated summaries with respect to the placement on the page (left or right). The annotators evaluated the summaries based on the following criteria:

[1] https://tinyurl.com/q5v9p5d

[2] https://tinyurl.com/y9k2f25b

[3] The main reason behind *ABNER* insufficiency is the presence of novel PubMed named entities that were not covered by any of the classes in the *ABNER* tool.

Non-Repetition and no factual Redundancy (Non-Re): There should not be redundancy in the factual information, and no repetition of sentences is allowed.

Coherence (Coh): Coherence means “continuity of sense”. The arguments have to be connected sensibly so that the reader can see consecutive sentences as being about one (or a related) concept.

Readability (Read): Consideration of general readability criteria such as good spelling, correct grammar, understandability, etc. in the summaries.

Informativeness, Overlap and Focus (IOF): How much information is covered by the summary. The goal is to find the common pieces of information via matching the same keywords (or key phrases), such as “Nematodes”, across the summary. For overlaps, annotators compare the keywords’ (or key-phrases) occurrence frequency and ensure the summaries are on the same topic.

Criteria	Mean (μ)	S.D. (σ)
Non-Re	7.19	0.755
Coh	6.87	0.705
Read	6.82	0.821
IOF	6.31	0.879

Table 2: Mean and Standard Deviation (SD) scores of human annotation on 50 summaries

The average scores and standard deviations are shown in Table 2. Annotators found that for readability, coherence, and non-repetitiveness, the quality of summaries is satisfactory. However, for informativeness and overlap, it is hard to evaluate summaries due to domain-specific technical terms.

ROUGE and Human Scores We calculated the Pearson’s correlation (Pearson, 1895) between ROUGE (Lin, 2004) scores (ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L)) in terms of precision, recall and F1 score with the human-evaluated scores.[4] Pearson’s correlation value (between -1 and $+1$) quantifies the degree to which quantitative and continuous variables are related to each other. The correlations are shown in Table 3.

ROUGE scores assume that a high-quality summary generated by a model should have common words and phrases with a gold-standard summary. However, this is not always true because (a) there can be semantically similar meaning (synonymous) word usage, and (b) there can be the usage of text paraphrases (similar information conveyed) with a little lexical overlap in the reference summary text. Therefore, merely considering lexical overlaps to evaluate summary quality is not sufficient. A high ROUGE score may indicate a good summary, but a low ROUGE score does not necessarily indicate a bad summary. Furthermore, while summarizing large documents, humans tend to utilize different paraphrasing/words to convey the same meaning in a shorter form. Several studies by Cohan and Goharian (2016); Dohare et al. (2017) argue that ROUGE is not an accurate estimator of the quality of a summary for scientific input, e.g., biomedical text. Hence, a weak correlation of ROUGE scores with human ratings on SUMPUBMED, as reported in Table 3, should not be a surprise.

4 Experiments

We have used the noun phrase version of SUMPUBMED in the abstractive summarization settings and the Hybrid version of SUMPUBMED in the extractive and the hybrid settings, i.e., (extractive + abstractive) summarizations. We split the dataset into train (93%), test (3%), and validation (4%) sets. Before training, we wrote a script that first tokenizes all input files and then forms the vocabulary and chunked files for the train, test, and validation sets. This step converts the input into a suitable format for the $seq2seq$ models.

Baseline Models We used extractive, abstractive, and hybrid (extractive + abstractive) summarization methods to evaluate SUMPUBMED. For abstractive summarization, we used two modifications of *Seq2Seq* models with attention: (a) *Basic Seq2Seq model (*seq2seq*)* which is a single layer bidirectional LSTM encoder and a single layer unidirectional LSTM decoder seq2seq model (See et al., 2017) with a copying mechanism (Gu et al., 2016) and attention (Nallapati et al., 2016). (b) *Seq2Seq model with Coverage (*seq2seq + *cov)* which is the earlier *seq2seq* model with a coverage mechanism (Mi et al., 2016) for penalizing phrase redundancy during generation. For extractive summarization, we used the unsupervised TextRank (Mihalcea and Tarau, 2004) method for sentence extraction. In hybrid approach, we first used extractive summarization (TextRank) and then apply abstractive summarization (*seq2seq* + cov) on the extracted text.[5]

[4] ROUGE-n is an n-gram similarity measure that computes uni/bi/trigram and higher n-gram overlaps. In R-L, L refers to the Longest Common Subsequence (LCS) overlap: a subsequence of matching words with the maximal length that is common in both texts with the order of words being preserved.

[5] We do not preprocess in the beginning like earlier, since the extractive summarization step makes the text length short

Criteria	Prec			Recall			F1		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Non-Re	-0.09	-0.06	-0.11	+0.02	-0.07	+0.007	+0.008	-0.05	+0.03
Coh	+0.05	-0.14	+0.05	-0.04	-0.25	-0.01	+0.02	-0.19	+0.06
Read	+0.19	+0.09	+0.20	+0.006	-0.03	+0.03	+0.12	+0.01	+0.13
IOF	-0.15	-0.18	-0.16	+0.12	0.08	+0.09	+0.06	-0.007	+0.12

Table 3: Pearson's correlation between ROUGE scores and human ratings on SumPubMed's noun-phrase version

Experimental Settings While decoding seq2seq models (for abstractive and hybrid models), we use a beam search (Medress et al., 1977) with a beam width of 4.[6] We also experimented with varying target summary lengths (i.e., the number of decoding steps) for seq2seq models. We report both seq2seq models with and without coverage results for comparison. We considered ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L)'s precision, recall, and F1 score for evaluation. Refer to the exact hyper-parameters in Appendix A.

5 Results and Analysis

Results on SumPubMed for abstractive methods, i.e., seq2seq models (with and without coverage), the extractive method of TextRank, and the hybrid approach, i.e., TextRank + seq2seq (with and without coverage) are shown in Tables 5, 6, and 7, respectively. We also evaluated the seq2seq models on news datasets (CNN/Daily Mail and DUC 2001) for comparison, as shown in Table 4.

Analysis: In all three approaches, abstractive in Table 5, extractive in Table 6 and hybrid in Table 7, we notice that the ROUGE Recall and F1-score increase, whereas precision decreases with the number of words (100 to 250) in the target summaries. The increase in Recall is expected as the chances of lexical overlap are more with larger generated summaries. Precision decreases because, with more words, the chances of non-covered words in the output summary also increase.

We notice in both Tables 5 and 7 that by adding the coverage (+cov) mechanism, the problem of repetition in summaries is solved to a great extent. The ROUGE scores also show improvement after applying coverage to pointer-generator networks. Thus, one can conclude that pointer generator networks effectively handle named entities and out-of-vocabulary words, and the coverage mechanism is useful to avoid repetitive generation, which is essential for scientific summarization.

In Table 8, we note that in terms of Precision (Pr), the abstractive approach shows the best results. However, the Recall (Re) of the extractive summarization model is always better than abstractive and hybrid approaches. Furthermore, the R-1 Re (ROUGE-1 Recall) and R-L Re (ROUGE-L Recall) for the hybrid models are approximately similar to the abstractive models. We also provide a few qualitative examples of summarization in section 7.

6 Related Work

Below, we provide the details of other summarization datasets:

News: CNN-Daily Mail has $92,000$ examples with documents of 30-sentence length with 4 corresponding human-written summaries of 50 words. DUC (Document Understanding Conference), another dataset, contains 500 documents (35.6 tokens on average) and summaries (10.4 tokens). Gigaword (Rush et al., 2015) has 31.4 document tokens and 8.3 summary tokens. Lastly, X-Sum (Extreme Summarization) (Narayan et al., 2018) contains 20-sentence (BBC articles) (431 words) and corresponding one-sentence (23 words) summaries.

Social Media: Webis-TLDR-17 Corpus (Völske et al., 2017) is a large-scale dataset of 3 million pairs of content and self-written summaries obtained from social media (Reddit). Webis-Snippet-20 Corpus (Chen et al., 2020) contains 10 million (webpage content and abstractive snippet) pairs and 3.5 million triples (query terms, abstractive snippets, etc.) for query-based abstractive snippet generation of web pages.

Scientific: Recently, Sharma et al. (2019) released a large dataset of 1.3 million of U.S. patent documents along with human written summaries. However, the closest datasets to SumPubMed are released by Cohan et al. (2018); Kedzie et al. (2018); Gidiotis and Tsoumakas (2019).

enough for the followup abstractive summarization.

[6] Beam search is a greedy technique which chooses the most likely token from all generated tokens at each step to obtain the best b sequences (the hyper-parameter b here represents the beam width). Beam search is shown to be better than generating the first sequence.

Data	Model	R-1			R-2			R-L		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
CNN	seq2seq	33.49	38.49	34.61	13.89	15.87	14.29	30.15	34.64	31.15
-DM	+cov	**38.59**	**41.10**	**38.53**	**16.84**	**17.83**	**16.75**	**35.56**	**37.81**	**35.48**
DUC	seq2seq	41.34	21.33	27.63	14.28	7.30	9.49	32.95	16.93	21.93
	+cov	**43.86**	**21.92**	**28.57**	**15.04**	**7.41**	**9.68**	**34.96**	**17.29**	**22.60**

Table 4: ROUGE scores on CNN-Dailymail (CNN-DM) and DUC 2001 dataset (DUC) using seq2seq models

Steps	Model	R-1			R-2			R-L		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
100	seq2seq	52.30	20.56	28.01	16.01	6.17	8.50	47.97	18.70	25.53
	+cov	**57.50**	22.66	31.04	**20.28**	7.74	10.73	**52.62**	20.56	28.23
150	seq2seq	48.88	27.10	32.81	15.18	8.35	10.18	44.64	24.56	29.81
	+cov	55.11	29.71	36.79	19.17	10.14	12.66	50.48	27.07	33.57
200	seq2seq	44.83	30.23	33.79	13.73	9.20	10.33	40.86	27.37	30.65
	+cov	52.86	33.84	39.21	18.25	11.52	13.43	48.47	30.88	35.84
250	seq2seq	41.18	31.84	33.00	12.80	9.79	10.22	37.68	28.89	30.03
	+cov	51.11	**36.24**	**40.13**	17.63	**12.39**	**13.77**	46.92	**33.13**	**36.73**

Table 5: ROUGE scores of noun-phrase SumPubMed version using a seq2seq model of varying decoding steps

Steps	R-1			R-2			R-L		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
150	**45.91**	31.69	36.82	**16.97**	11.09	13.12	39.12	26.91	28.84
200	42.81	36.03	38.44	15.71	13.31	14.10	36.60	30.73	31.48
250	40.51	**39.59**	**39.33**	14.81	**15.30**	**14.72**	34.83	33.98	**34.83**

Table 6: Results for TextRank an Extractive Summarization approach on hybrid version of the SumPubMed.

Steps	Model	R-1			R-2			R-L		
		Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
100	seq2seq	50.32	21.09	28.45	12.66	5.14	7.04	46.58	19.40	26.23
	+cov	**56.07**	27.42	30.69	**16.65**	6.47	8.95	**51.87**	20.62	28.27
150	seq2seq	45.01	25.50	30.99	11.14	6.21	7.59	41.43	23.35	28.42
	+cov	52.23	29.11	35.62	15.44	8.45	10.42	48.35	26.81	32.86
200	seq2seq	40.55	28.46	31.56	9.93	6.93	7.70	37.21	25.98	28.86
	+cov	47.82	33.37	37.28	14.01	9.68	10.84	44.29	30.80	34.44
250	seq2seq	35.80	30.88	30.61	9.14	7.67	7.66	32.67	27.95	27.80
	+cov	43.82	**36.16**	**37.33**	12.77	**10.49**	**10.85**	40.55	**33.37**	**34.49**

Table 7: ROUGE scores on hybrid version of the SumPubMed using Hybrid model: TextRank + seq2seq models

Model	R-1			R-2			R-L		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
Abstractive	**51.11**	36.24	**40.13**	**17.63**	12.39	**13.77**	**46.92**	33.13	**36.73**
Extractive	40.51	**39.59**	39.33	14.81	**15.30**	14.72	34.83	**33.98**	32.82
Hybrid Model	43.82	36.16	37.33	12.77	10.49	10.85	40.55	33.37	34.49

Table 8: ROUGE comparison on SumPubMed. seq2seq abstractive methods' target summary is of 250 words

Comparison with SumPubMed: News datasets' summary is located at the top of the article for most examples. Social media datasets lack the scientific aspect, i.e., complex domain-specific vocabulary and non-localized distributed information of SumPubMed. Other works on the scientific datasets are by Cohan et al. (2018); Kedzie et al. (2018); Gidiotis and Tsoumakas (2019). The closest work to our approach is the PubMed dataset by Cohan et al. (2018). However, unlike SumPubMed, (a) no extensive preprocessing pipeline was applied to clean the text (b) a single version is released compared with SumPubMed's several versions with distinct properties (varying summary lengths, article lengths, and vocabulary sizes), (c) only level-1 section headings instead of the whole PubMed document are used, and (d) there is a lack of human evaluation to assess data quality.

7 Example of Summarization on SumPubMed

Here we provide an representative examples of actual summary. Repetitiveness i.e. factual redundancy is shown with the highlighted text.

7.1 Abstractive Summarization on SUMPUBMED

We see that factual redundancy and repetitiveness in the generated summaries with pointer-generation which is removed by applying coverage. We also observe that repetitiveness is removed by using the coverage mechanism.

Reference Summary the origin of these genes has been attributed to horizontal gene transfer from bacteria, although there still is a lot of uncertainty about the origin and structure of the ancestral ghf <dig> ppn endoglucanase. our data confirm a close relationship between pratylenchus spp. furthermore, based on gene structure data, we inferred a model for the evolution of the ghf <dig> endoglucanase gene structure in plantparasitic nematodes. our evolutionary model for the gene structure in ppn ghf <dig> endoglucanases implies the occurrence of an early duplication event, and more recent gene duplications at genus or species level. the latter one is the first gene isolated from a ppn of a different superfamily -LRB- sphaerularioidea -RRB-; all previously known nematode endoglucanases belong to the superfamily tylenchoidea -LRB- order rhabditida -RRB-. no statistical incongruence between the phylogenetic trees deduced from the catalytic domain and the cbm <dig> was found, which could suggest that both domains have evolved together. and the root knot nematodes, while some radopholus similis endoglucanases are more similar to cyst nematode genes. two new endoglucanases from the migratory nematodes pratylenchus coffeae and ditylenchus africanus were included in this study. phylogenetic analyses were conducted with the ppn ghf <dig> endoglucanases and homologous endoglucanases from bacterial and other eukaryotic lineages such as beetles, fungi and plants. previous studies on the evolution of these genes have focused primarily on data of sedentary nematodes, while in this study, extra data from migratory nematodes were included. it is not clear whether this ancestral endoglucanase consisted of the whole gene cassette, containing a catalytic domain and a carbohydratebinding module -LRB- cbm, type <dig> in ppn and bacteria -RRB- or only of the catalytic domain while the cbm <dig> was retrieved by domain shuffling later in evolution. the ghf <dig> catalytic domain and the cbm <dig> rather than that it evolved by domain shuffling. endo <dig> betaglucanases or cellulases from the glycosyl hydrolase family <dig> -LRB- ghf5 -RRB- have been found in numerous bacteria and fungi, and recently also in higher eukaryotes, particularly in plantparasitic nematodes -LRB- ppn -RRB-.

Summary from seq2seq although the hox cell line with peculiar properties of horizontal gene transfer -LRB- hgt -RRB- has been proposed to delineate the function of ghf <dig> endoglucanases. the gene structure of the ppn ghf <dig> endoglucanases is a matter of debate, in this study no support for an early exon/domain shuffling event after the independent gain of a catalytic domain. the exon/intron gene structure of the ppn ghf <dig> endoglucanases was compared with the multiple protein sequence alignment. the gene structure of the ppn ghf <dig> endoglucanases was compared with the multiple protein sequence alignment. **the gene structure of the ppn ghf <dig> endoglucanases was compared with the multiple protein sequence alignment. the gene structure of the ppn ghf <dig> endoglucanases was compared with the multiple protein sequence alignment. the gene structure of the ppn ghf <dig> endoglucanases was compared with the multiple protein sequence alignment.** the evolution of the gene structure of the ppn ghf <dig> gene families: the number of members from an ancient or early eukaryotic ancestral gene is associated with the expansion of members from an ancient or early eukaryotic ancestral gene. in this study, we.

Summary from seq2seq with coverage the expansion of horizontal gene transfer -LRB- hgt -RRB- events in horizontal gene transfer -LRB- hgt -RRB- has been proposed to explain the origin of ghf <dig> endoglucanases in the nematode kingdom. while the ppn ghf <dig> endoglucanases has a close relationship to the root knot nematodes. in order to have a broader overview of the endoglucanase evolution in the infraorder tylenchomorpha, the gene structure of six additional genes was incorporated in our study. the ppn ghf <dig> gene family is associated with the expansion of the ppn ghf <dig> gene family bordered by intron <dig> and intron <dig> although 1 - <dig> symmetrical domains are suggested to be frequently associated with domain shuffling events in the evolution of paralogous gene families: the evolution of the ppn indicate a history of recent duplication events for which little information is available. our model implies that the divergence of the gene structure of the ppn ghf <dig> gene family is notably dynamic, and this evolution involves more intron gains than losses in the order rhabditida -LRB- infraorder tylenchomorpha -RRB-, which is part of one of the three evolutionary independent plantparasitic nematode clades. our results demonstrate that the conserved gene structure of the ppn ghf <dig> endoglucanases and the observation of some sequence conservation in the evolution of the plantparasitic bacteria and nematodes. our results suggest that the evolution of the ghf <dig> gene family is a major consequence of the evolution of.

7.2 Extractive Summarization on SumPubMed

TextRank produces a purely extractive summary. But we see that it is able to identify the relevant sentences. The content overlap between the reference and generated extractive summary is adequate.

Reference Summary **to find out the different ovarian activity and follicle recruitment with mirnamediated posttranscriptional regulation, the small rnas expressed pattern in the ovarian tissues of multiple and uniparous anhui white goats during follicular phase was analyzed using solexa sequencing data.** <dig> mirnas coexpressed, <dig> and <dig> mirnas specifically expressed **in the ovaries of multiple and uniparous goats during follicular phase** were identified. in the present study, the different expression of mirnas in the ovaries of multiple and uniparous goats during follicular phase were characterized and investigated using deep sequencing technology. rt-pcr was applied to detect the expression level of <dig> randomly selected mirnas in multiple and uniparous hircine ovaries, and the results were consistent with the solexa sequencing data. micrornas play critical roles in almost all ovarian biological processes, including folliculogenesis, follicle development, follicle atresia, luteal development and regression. the result will help to further understand the role of mirnas in kidding rate regulation and also may help to identify mirnas which could be potentially used to increase hircine ovulation rate and kidding rate in the future. the <dig> most highly expressed mirnas in the multiple library were also the highest expressed in the uniparous library, and there were no significantly different between each other. **the highest specific expressed mirna in the multiple library was mir29c, and the one in the uniparous library was mir<dig>** <dig> novel mirnas were predicted in total. superior kidding rate is an important economic trait in production of meat goat, and ovulation rate is the precondition of kidding rate. go annotation and kegg pathway analyses were implemented on target genes of all mirna in two libraries.

Extracted Summary **in order to identify differentially expressed mirna during follicular phase in the ovaries of multiple and uniparous anhui white goats, two small rna libraries were constructed by solexa sequencing.** for all mirnas target genes of multiple and uniparous goats in the ovaries during follicular phase, there were <dig> and <dig> target genes mapped to the go terms of cellular component. the expression levels of <dig> randomly selected mirnas were verified in the ovaries of multiple and uniparous goats during follicular phase using rt-pcr. in this study, we sequenced the small rnas **in the ovarian tissues of multiple and uniparous anhui white goats during follicular phase** by illumina solexa technology, then analyzed the differentially expressed mirnas, predicted novel mirnas, and made go enrichment and kegg pathway analysis of target genes in two mirna libraries. in ovaries between multiple and uniparous goats of follicular phase, <dig> novel mirnas were predicted in total, which is distinctly more than the amount predicted in our previous study implemented by our team workers, zhang et al. **the highest specific expressed mirna in multiple library was mir29c, and the one in uniparous library was mir<dig>** as aligning the clean reads to the mirna precursor/mature mirnas of all animals in the mirbase <dig> database, and obtained mirna with no specified species. rt-pcr was carried out to analyze the expression of <dig> randomly selected mirnas in multiple and uniparous hircine ovaries during follicular phase, and the results were consistent with the solexa sequencing data.

8 Conclusion

We created a scientific summarization dataset, SumPubMed, to study the task of scientific summarization. We showed how summarization on news articles is not reliable, thus indicating the need for SumPubMed. We also performed a human evaluation on SumPubMed based on the four crucial dimensions of readability, coherence, non-repetition, and informativeness. We built several extractive, abstractive, and hybrid baseline models to examine SumPubMed.

References

Wei-Fan Chen, Shahbaz Syed, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Abstractive Snippet Generation. In *Web Conference (WWW 2020)*, pages 1309–1319. ACM.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 615–621.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813.

Shibhansh Dohare, Harish Karnick, and Vivek Gupta. 2017. Text summarization using abstract meaning representation. *arXiv preprint arXiv:1706.01678*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Annemarie Friedrich, Marina Valeeva, and Alexis Palmer. 2014. LQVSumm: A corpus of linguistic quality violations in multi-document summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1591–1599, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alexios Gidiotis and Grigorios Tsoumakas. 2019. Structured summarization of academic publications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 636–645. Springer.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mark F. Medress, Franklin S Cooper, Jim W. Forgie, CC Green, Dennis H. Klatt, Michael H. O'Malley, Edward P Neuburg, Allen Newell, DR Reddy, B Ritea, et al. 1977. Speech understanding systems: Report of a steering committee. *Artificial Intelligence*, 9(3):307–316.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

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.

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.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.