Exploring Content Selection in Summarization of Novel Chapters

Faisal Ladhak[1]*, Bryan Li[2]*, Yaser Al-Onaizan[3], Kathleen McKeown[1,3]

[1]Columbia University, [2]University of Pennsylvania, [3]Amazon AI

`faisal.ladhak@columbia.edu, bryanli@seas.upenn.edu,`
`onaizan@amazon.com, kathy@cs.columbia.edu`

Abstract

We present a new summarization task, generating summaries of novel chapters using summary/chapter pairs from online study guides. This is a harder task than the news summarization task, given the chapter length as well as the extreme paraphrasing and generalization found in the summaries. We focus on extractive summarization, which requires the creation of a gold-standard set of extractive summaries. We present a new metric for aligning reference summary sentences with chapter sentences to create gold extracts and also experiment with different alignment methods. Our experiments demonstrate significant improvement over prior alignment approaches for our task as shown through automatic metrics and a crowd-sourced pyramid analysis.

1 Introduction

When picking up a novel one is reading, it would be helpful to be reminded of what happened last. To address this need, we develop an approach to generate extractive summaries of novel chapters. This is much harder than the news summarization tasks on which most of the summarization field (e.g., (Cheng and Lapata, 2016; Grusky et al., 2018; Paulus et al., 2017)) focuses; chapters are on average seven times longer than news articles. There is no one-to-one correspondence between summary and chapter sentences, and the summaries in our dataset use extensive paraphrasing, while news summaries copy most of their information from the words used in the article.

We focus on the task of content selection, taking an initial, extractive summarization approach given the task difficulty.[1] As the reference sum-

maries are abstractive, training our model requires creating a gold-standard set of extractive summaries. We present a new approach for aligning chapter sentences with the abstractive summary sentences, incorporating weighting to ROUGE (Lin, 2004) and METEOR (Lavie and Denkowski, 2009) metrics to enable the alignment of salient words between them. We also experiment with BERT (Devlin et al., 2018) alignment.

We use a stable matching algorithm to select the best alignments, and show that enforcing one-to-one alignments between reference summary sentences and chapter sentences is the best alignment method of those used in earlier work.

We obtain a dataset of summaries from five study guide websites paired with chapter text from Project Gutenberg. Our dataset consists of 4,383 unique chapters, each of which is paired with two to five human-written summaries.

We experiment with generating summaries using our new alignment method within three models that have been developed for single document news summarization (Chen and Bansal, 2018; Kedzie et al., 2018; Nallapati et al., 2017). Our evaluation using automated metrics as well as a crowd-sourced pyramid evaluation shows that using the new alignment method produces significantly better results than prior work.

We also experiment with extraction at different levels of granularity, hypothesizing that extracting constituents will work better than extracting sentences, since summary sentences often combine information from several different chapter sentences. Here, our results are mixed and we offer an explanation for why this might be the case.

Our contributions include a new, challenging summarization task, experimentation that reveals potential problems with previous methods for creating extracts, and an improved method for creating gold standard extracts.

---

*Equal contribution. Work done while at Amazon.

[1]We tried two abstractive models (Chen and Bansal, 2018; Liu and Lapata, 2019) but ROUGE was low and the output was poor with many repetitions and hallucinations.

2 Related Work

Relatively little work has been done in summarization of novels, but early work (Mihalcea and Ceylan, 2007) provided a dataset of novel/summary pairs drawn from CliffsNotes and GradeSaver and developed an unsupervised system based on Meade (Radev et al., 2001) and TextRank (Mihalcea and Tarau, 2004) that showed promise. More recently, Zhang et al. (2019) developed an approach for summarizing characters within a novel. We hypothesize that our proposed task is more feasible than summarizing the full novel.

Previous work has summarized documents using Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to extract elementary discourse units (EDUs) for compression and more content-packed summaries (Daumé III and Marcu, 2002; Li et al., 2016; Arumae et al., 2019). Some abstractive neural methods propose attention to focus on phrases within a sentence to extract (Gehrmann et al., 2018). Fully abstractive methods are not yet appropriate for our task due to extensive paraphrasing and generalization.

While previous work on semantic textual similarity is relevant to the problem of finding alignments between chapter and summary text, the data available (Cer et al., 2017; Dolan and Brockett, 2005) is not suitable for our domain, and the alignments we generated from this data were of a poorer quality than the other methods in our paper.

3 Data

We collect summary-chapter pairs from five online study guides: BarronsBookNotes (BB), BookWolf (BW), CliffsNotes (CN), GradeSaver (GS) and NovelGuide (NG).[2] We select summaries from these sources for which the complete novel text can be found on Project Gutenberg.

Our initial dataset, for summaries with two or more sources, includes 9,560 chapter/summary pairs for 4,383 chapters drawn from 79 unique books. As our analysis shows a very long tail, two rounds of filtering were applied. First, we remove reference texts with >700 sentences, as these are too large to fit into mini-batches ($\sim$10% of data). Second, we remove summaries with a compression ratio of <2.0, as such wordy summaries often

---

[2]We do not have the rights to redistribute the data. To allow others to replicate the dataset, we provide a list of novel chapters we used at https://github.com/manestay/novel-chapter-dataset

Summary Source	Mean (stdev)	Median	Total #
CN	442 (369)	347	1,053
BB	517 (388)	429	1,000
GS	312 (311)	230	1,983
BW	276 (232)	214	182
NG	334 (302)	244	2,070
All Sources	373 (339)	279	6,288
Chapter Text	5,165 (3,737)	4,122	6,288

Table 1: **Train Split Statistics**: World count statistics with total number for summaries and chapter text.

contain a lot of commentary (i.e. phrases that have no correspondence in the chapter, $\sim$5%).

This results in 8,088 chapter/summary pairs, and we randomly assign each book to train, development and test splits (6,288/938/862 pairs respectively). After filtering, chapters are on average seven times longer than news articles from CNN/Dailymail (5,165 vs 761 words), and chapter summaries are eight times longer than news summaries (372 vs 46 words).

Train split statistics are given in Table 1. These statistics reveal the large variation in length. Furthermore, we calculate *word overlap*, the proportion of vocabulary that overlaps between the summary and chapter. For novels, this is 33.7%; for CNN/DailyMail news, this is 68.7%. This indicates the large amount of paraphrasing in the chapter summaries in relation to the original chapter.

In Figure 1, we show the first three sentences of a reference summary for Chapter 11, *The Awakening* which is paraphrased from several, non-consecutive chapter sentences shown near the bottom of the figure. We also show a portion of the summaries from two other sources which convey the same content and illustrate the extreme level of paraphrasing as well as differences in detail. We show the full chapter and three full reference summaries in Appendix A.2.

4 Alignment Experiments

To train models for content selection, we need saliency labels for each chapter segment that serve as proxy extract labels, since there are no gold extracts. In news summarization, these are typically produced by aligning reference summaries to the best matching sentences from the news article. Here, we align the reference summary sentences with sentences from the chapter. We address two questions for aligning chapter and summary sentences to generate gold standard extracts:

GS: In this chapter Mr. and Mrs. Pontellier participate in a battle of wills. When Mr. Pontellier gets back from the beach, he asks his wife to come inside. She tells him not to wait for her, at which point he becomes irritable and more forcefully tells her to come inside.

NG: Mr. Pontellier is surprised to find Edna still outside when he returns from escorting Madame Lebrun home. ... although he asks her to come in to the house with him, she refuses, and remains outside, exercising her own will.

BW: Leonce urges Edna to go to bed, but she is still exhilarated and decides to stay outside in the hammock...

Chapter sentences: He had walked up with Madame Lebrun and left her at the house. "Do you know it is past one o'clock? Come on," and he mounted the steps and went into their room. "Don't wait for me," she answered. "You will take cold out there," he said, irritably. "What folly is this? Why don't you come in?"

Figure 1: Portions of three reference summaries for *The Awakening*, Chapter 11 by Kate Chopin, along with chapter sentences they summarize.

1) Which similarity metric works best for alignment (Section 4.1)? and 2) Which alignment method works best (Section 4.2)?

4.1 Similarity Metrics

ROUGE is commonly used as a similarity metric to align the input document and the gold standard summary to produce gold extracts (Chen and Bansal, 2018; Nallapati et al., 2017; Kedzie et al., 2018). One drawback to using ROUGE as a similarity metric is that it weights all words equally. We want to, instead, assign a higher weight for the salient words of a particular sentence.

To achieve this, we incorporate a smooth inverse frequency weighting scheme (Arora et al., 2017) to compute word weights. The weight of a given word is computed as follows:

$$W(w_i) = \frac{\alpha}{\alpha + p(w_i)} \quad (1)$$

where $p(w_i)$ is estimated from the chapter text and α is a smoothing parameter (here $\alpha = 1e^{-3}$). N-gram and Longest Common Subsequence (LCS) weights are derived by summing the weights of each of the individual words in the N-gram/LCS. We take the average of ROUGE-1, 2, L using this weighting scheme as the metric for generating extracts, *R-wtd*, incorporating a stemmer to match morphological variants (Porter, 1980).

Similarity Metrics Results: We compare R-wtd against ROUGE-L (Chen and Bansal, 2018) (R-L), and ROUGE-1, with stop-word removal and stemming (Kedzie et al., 2018) (R-1), for sentence alignment. To incorporate paraphrasing, we average METEOR (Banerjee and Lavie, 2005) scores with ROUGE-1,2,L for both un-weighted (RM) and weighted scores (RM-wtd). Given the recent success of large, pre-trained language models for downstream NLP tasks, we also experiment with BERT (Devlin et al., 2019) to compute alignment, using cosine similarity between averaged chapter segment and summary segment vectors. We compare the generated gold extracts using R-L F1 against reference summaries, to determine a shortlist for human evaluation (to save costs).

For the human evaluation, we ask crowd workers to measure content overlap between the generated alignments, and the reference summary, on a subset of the validation data. For each summary reference, they are shown a generated alignment and asked to indicate whether it conveys each of up to 12 summary reference sentences. An example task is shown in Appendix Figure 7. We then compute precision and recall based on the number of summary sentences conveyed in the extract.

Table 2 shows that humans prefer alignments generated using R-wtd by a significant margin.[3] Sample alignments generated by R-wtd in comparison to the baseline are shown in Figure 2.

Method	RM	R-wtd	RM-wtd	R-1	R-L	BERT
R-L F1	**41.2**	**40.6**	39.3	37.1	35.1	35.4
H-F1	33.7	**44.8**	38.8	–	–	–

Table 2: ROUGE-L F1, and crowd-sourced F1 scores (H-F1) for content overlap.

4.2 Alignment Methods

Some previous work in news summarization has focused on iteratively picking the best article sentence with respect to the summary, in order to get the gold extracts (Nallapati et al., 2017; Kedzie et al., 2018), using ROUGE between the set of selected sentences and the target summary. In contrast, others have focused on picking the best article sentence with respect to each sentence in the summary (Chen and Bansal, 2018). We investigate which approach yields better alignments. We refer

---

[3] We suspect incorporating METEOR by averaging didn't work because the scale is different from ROUGE scores.

to the former method as summary-level alignment and the latter method as sentence-level alignment.

For sentence-level alignment, we note that the problem of finding optimal alignments is similar to a stable matching problem. We wish to find a set of alignments such that there exists no chapter segment a and summary segment x where both a and x would prefer to be aligned with each other over their current alignment match. We compute alignments based on the Gale-Shapley algorithm (1962) for stable matching and compare it with the greedy approach from prior work (Chen and Bansal, 2018).

For summary-level alignment (Nallapati et al., 2017; Kedzie et al., 2018), we compare two variants: selecting sentences until we reach the reference word count (WL summary), and selecting sentences until the ROUGE score no longer increases (WS summary).

Crowd-sourced evaluation results (Table 3) show that sentence-level stable matching is significantly better. We use this in the remainder of this work. These differences in alignments affect earlier claims about the performance of summarization systems, as they were not measured, yet have a significant impact.[4]

Method	P	R	F1
Greedy Sent	48.4	48.7	48.5
Stable Sent	**52.8**	**52.6**	**52.7**
WL summary	34.5	36.6	36.7
WS summary	42.7	36.6	38.0

Table 3: Crowd sourced evaluation on content overlap for summary vs. sentence level on validation set.

> **Ref summary:** He says he will, as soon as he has finished his last cigar.
> **R-L greedy:** "You will take cold out there," he said, irritably.
> **R-L stable:** He drew up the rocker, hoisted his slippered feet on the rail, and proceeded to smoke a cigar.
> **R-wtd stable:** "Just as soon as I have finished my cigar."

Figure 2: A reference summary sentence and its alignments. R-L greedy and R-L stable are incorrect because they weight words equally (e.g. said, cigar, '.').

---

[4]Bold text indicates statistical significance with $p < 0.05$.

5 Summarization Experiments

In order to assess how alignments impact summarization, we train three extractive systems – hierarchical CNN-LSTM extractor (Chen and Bansal, 2018) (CB), seq2seq with attention (Kedzie et al., 2018) (K), and RNN (Nallapati et al., 2017) (N). The target word length of generated summaries is based on the average summary length of similarly long chapters from the training set.[5]

We also experiment with aligning and extracting at the constituent level,[6] given our observation during data analysis that summary sentences are often drawn from two different chapter sentences. We create syntactic constituents by taking sub-trees from constituent parse trees for each sentence (Manning et al., 2014) rooted with S-tags. To ensure that constituents are long enough to be meaningful, we take the longest S-tag when one S-tag is embedded within others (see Appendix A.5).

Summary quality is evaluated on F1 scores for R-{1,2,L}, and METEOR. Each chapter has 2-5 reference summaries and we evaluate the generated summary against all the reference summaries. Part of a generated summary of extracted constituents for Chapter 11, *The Awakening*, is shown in Figure 3. The full generated summaries for this chapter (both extracted constituents and extracted sentences) are shown in Appendix A.2.

> **Generated Summary:** |I thought I should find you in bed , " ||said her husband , |when he discovered her |lying there . |He had walked up with Madame Lebrun and left her at the house . ||She heard him moving about the room ; |every sound indicating impatience and irritation .|

Figure 3: System generated summary, extracted constituents in teal, and separated by |.

5.1 Results

We compare our method for generating extractive targets (ROUGE weighted, with stable matching at the sentence level) against the baseline method for generating extractive targets for each of the systems. Table 4 shows three rows for each summarization system: using the original target summary labels, and using either constituent or sentence segments. We see our proposed alignment method performs significantly better for all mod-

---

[5]We do so by binning chapters into 10 quantiles by length.
[6]Prior work has used EDUs, but automated parsers such as (Ji and Eisenstein, 2014) perform poorly in this domain.

Model	Seg	Method	R-1	R-2	R-L	METEOR
	sent	baseline	33.1	5.5	30.0	13.9
CB	sent	R-wtd	**35.8**	**6.9**	33.4	**15.2**
	const	R-wtd	**36.2**	**6.9**	**35.4**	15.2
	sent	baseline	34.3	6.4	31.6	14.6
K	sent	R-wtd	**35.6**	**6.9**	33.2	**15.0**
	const	R-wtd	**36.2**	**6.9**	**35.2**	15.1
	sent	baseline	34.6	6.4	31.9	14.6
N	sent	R-wtd	**35.7**	**7.0**	33.3	**15.1**
	const	R-wtd	**35.9**	**7.0**	**35.2**	15.0

Table 4: ROUGE-F1, METEOR for generated summaries. "Baseline" is the method used for that model.

els. ROUGE-L in particular increases 10% to 18% relatively over the baselines. Moreover, it would seem at first glance that the K and N baseline models perform better than the CB baseline, however this difference has nothing to do with the architecture choice. When we use our extractive targets, all three models perform similarly, suggesting that the differences are mainly due to small, but important, differences in their methods for generating extractive targets.

Human Evaluation: Given questions about the reliability of ROUGE (Novikova et al., 2017; Chaganty et al., 2018), we perform human evaluation to assess which system is best at content selection. We use a lightweight, sampling based approach for pyramid analysis that relies on crowd-sourcing, proposed by Shapira et al. (2019), and correlates well with the original pyramid method (Nenkova et al., 2007). We ask the crowd workers to indicate which of the sampled reference summary content units are conveyed in the generated summary.[7]

We evaluated our best system + alignment on extraction of sentences and of constituents (CB R-wtd), along with a baseline system (CB K-align),[8] using the crowd-sourced pyramid evaluation method. To produce readable summaries for extracted constituents, each extracted constituent is included along with the context of the containing sentence (black text in Figure 3). We find that CB Sent R-wtd has significantly higher content overlap with reference summaries in Table 5.

6 Discussion and Conclusion

We present a new challenging task for summarization of novel chapters. We show that *sentence-*

---

[7]See the screen shot in Appendix A.4

[8]We use the best baseline alignment, Kedzie et al. (2018) with the CB model to keep model choice consistent.

System	Pyramid Score
CB K-align	17.9
CB Sent R-wtd	**18.9**
CB Const R-wtd	18.1

Table 5: Crowd-sourced Pyramid Evaluation.

level, stable-matched alignment is better than the summary-level alignment used in previous work and our proposed *R-wtd* method for creating gold extracts is shown to be better than other similarity metrics. The resulting system is the first step towards addressing this task.

While both human evaluation and automated metrics concur that summaries produced with our new alignment approach outperform previous approaches, they contradict on the question of whether extraction is better at the constituent or the sentence level. We hypothesize that because we use ROUGE to score summaries of extracted constituents without context, the selected content is packed into the word budget; there is no potentially irrelevant context to count against the system. In contrast, we do include sentence context in the pyramid evaluation in order to make the summaries readable for humans and thus, fewer constituents make it into the generated summary for the human evaluation. This could account for the increased score on automated metrics.

It is also possible that smaller constituents can be matched to phrases within the summary with metrics such as ROUGE, when they actually should not have counted. In future work, we plan to experiment more with this, examining how we can combine constituents to make fluent sentences without including potentially irrelevant context.

We would also like to further experiment with abstractive summarization to re-examine whether large, pre-trained language models (Liu and Lapata, 2019) can be improved for our domain. We suspect these models are problematic for our documents because they are, on average, an order of magnitude larger than what was used for pre-training the language model (512 tokens). Another issue is that the pre-trained language models are very large and take up a substantial amount of GPU memory, which limits how long the input document can be. While truncation of a document may not hurt performance in the news domain due to the heavy lede bias, in our domain, truncation can hurt the performance of the summarizer.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Kristjan Arumae, Parminder Bhatia, and Fei Liu. 2019. Towards annotating and creating summary highlights at sub-sentence level. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 64–69, Hong Kong, China. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Melbourne, Australia. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686.

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.

Hal Daumé III and Daniel Marcu. 2002. A noisy-channel model for document compression. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 449–456. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

David Gale and Lloyd S Shapley. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 13–24.

Chris Kedzie, Kathleen McKeown, and Hal Daume 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.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3728–3738, Hong Kong, China. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text: Interdisciplinary Journal for the Study of Discourse*, 8(3):243281.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jennỹ Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

R. Mihalcea and P. Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

Rada Mihalcea and Hakan Ceylan. 2007. Explorations in automatic book summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 380–389, Prague, Czech Republic. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2).

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Dragomir Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multi-document summarization using MEAD. In *First Document Understanding Conference*, New Orleans, LA.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. 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 682–687, Minneapolis, Minnesota. Association for Computational Linguistics.

Weiwei Zhang, Jackie Chi Kit Cheung, and Joel Oren. 2019. Generating character descriptions for automatic summarization of fiction. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7476–7483.

A Appendix

A.2 Example Chapter and Summaries

We show the full text of Chapter 11, *The Awakening* by Kate Chopin in Figure 4. We show three reference summaries in Figure 5, and two generated summaries using our best alignment method in Figure 6. While there are differences in length and level of detail, there are also clearly similarities in covered content.

A.3 Target Word Length for Summaries

The target word length for generated summaries is a function of the input chapter word count ($wc_{chapter}$).

We divide the train set into 10 quantiles, and in each quantile (or bin), associate it to the mean *compression ratio* (CR):

$$CR = \frac{wc_{chapter}}{wc_{ref\_summ}} \quad (2)$$

$$CR_{quantile} = \frac{1}{n}\sum_{i=1}^{n} CR_i \quad (3)$$

Where wc_{ref_summ} is the word count of the reference summary, and CR_i is the compression ratio of the i-th quantile item.

The target word length for the generated summary (wc_{gen_summ}) is given by:

$$wc_{gen\_summ} = \frac{1}{CR_{quantile}} * wc_{chapter} \quad (4)$$

Generated summaries are created by extracting segments with the highest model probability until this budget is reached (without truncation). Oracle summaries also use this target word length, but may be shorter if the original summary had few segments (as we extract one chapter segment for each summary segment).

Quantile	Min wc	Max wc	CR
1	44	1,232	6.67
2	1,233	1,711	9.09
3	1,712	2,174	9.09
4	2,175	2,758	10.00
5	2,579	3,361	11.11
6	3,362	4,165	12.5
7	4,166	5,374	14.29
8	5,375	7,762	14.29
9	7,763	13,028	16.67
10	13,029	70,436	20

Table 6: **Quantiles**: For each quantile (bin), we show its max and min word words, and its compression ratio.

A.4 SCU Evaluation Task Setup

To obtain the *distractors*, we sample 2 SCUs from different chapters from the same book. We insert one of them, the positive distractor, into the generated summary, as well as into the list of statements, so it will always be correct. We insert the other, the negative distractor, only into the list of statements, so it will always be incorrect.

A.5 Constituent Extraction algorithm

Algorithm 1 extracts subtrees from a constituent parse tree. These subtrees are *constituents*, and break down sentences into meaningful spans of text. Constituents are one of

1. A relative clause
2. The highest level S or SBAR node in its subtree with (NP, VP) children
3. The highest level VP node above 2)
4. The remaining nodes in the tree that were not extracted with 1), 2) or 3)

"What are you doing out here, Edna? I thought I should find you in bed," said her husband, when he discovered her lying there. He had walked up with Madame Lebrun and left her at the house. His wife did not reply.

"Are you asleep?" he asked, bending down close to look at her.

"No." Her eyes gleamed bright and intense, with no sleepy shadows, as they looked into his.

"Do you know it is past one o'clock? Come on," and he mounted the steps and went into their room.

"Edna!" called Mr. Pontellier from within, after a few moments had gone by.

"Don't wait for me," she answered. He thrust his head through the door.

"You will take cold out there," he said, irritably. "What folly is this? Why don't you come in?"

"It isn't cold; I have my shawl."

"The mosquitoes will devour you."

"There are no mosquitoes."

She heard him moving about the room; every sound indicating impatience and irritation. Another time she would have gone in at his request. She would, through habit, have yielded to his desire; not with any sense of submission or obedience to his compelling wishes, but unthinkingly, as we walk, move, sit, stand, go through the daily treadmill of the life which has been portioned out to us.

"Edna, dear, are you not coming in soon?" he asked again, this time fondly, with a note of entreaty.

"No; I am going to stay out here."

"This is more than folly," he blurted out. "I can't permit you to stay out there all night. You must come in the house instantly."

With a writhing motion she settled herself more securely in the hammock. She perceived that her will had blazed up, stubborn and resistant. She could not at that moment have done other than denied and resisted. She wondered if her husband had ever spoken to her like that before, and if she had submitted to his command. Of course she had; she remembered that she had. But she could not realize why or how she should have yielded, feeling as she then did.

"Leonce, go to bed," she said, "I mean to stay out here. I don't wish to go in, and I don't intend to. Don't speak to me like that again; I shall not answer you."

Mr. Pontellier had prepared for bed, but he slipped on an extra garment. He opened a bottle of wine, of which he kept a small and select supply in a buffet of his own. He drank a glass of the wine and went out on the gallery and offered a glass to his wife. She did not wish any. He drew up the rocker, hoisted his slippered feet on the rail, and proceeded to smoke a cigar. He smoked two cigars; then he went inside and drank another glass of wine. Mrs. Pontellier again declined to accept a glass when it was offered to her. Mr. Pontellier once more seated himself with elevated feet, and after a reasonable interval of time smoked some more cigars.

Edna began to feel like one who awakens gradually out of a dream, a delicious, grotesque, impossible dream, to feel again the realities pressing into her soul. The physical need for sleep began to overtake her; the exuberance which had sustained and exalted her spirit left her helpless and yielding to the conditions which crowded her in.

The stillest hour of the night had come, the hour before dawn, when the world seems to hold its breath. The moon hung low, and had turned from silver to copper in the sleeping sky. The old owl no longer hooted, and the water-oaks had ceased to moan as they bent their heads.

Edna arose, cramped from lying so long and still in the hammock. She tottered up the steps, clutching feebly at the post before passing into the house.

"Are you coming in, Leonce?" she asked, turning her face toward her husband.

"Yes, dear," he answered, with a glance following a misty puff of smoke. "Just as soon as I have finished my cigar."

Figure 4: Full chapter text. Note that this is short at 847 words, as the median chapter length is 3168 words.

BookWolf summary: Léonce urges Edna to go to bed, but she is still exhilarated and decides to stay outside in the hammock. Léonce stays up with her and smokes his cigars. Edna feels defiant towards her husband and resents his control over her life. Eventually tiredness overcomes Edna and she goes to bed.

GradeSaver summary: In this chapter Mr. and Mrs. Pontellier participate in a battle of wills. When Mr. Pontellier gets back from the beach, he asks his wife to come inside. She tells him not to wait for her, at which point he becomes irritable and more forcefully tells her to come inside. Mrs. Pontellier resolves not to go in and thinks about how, on another occasion, she would have just done what her husband asked, simply because of inertia. Feeling stubborn and strong, she realizes that she had never taken such a stand against her husband before.
Mr. Pontellier then decides to join her outside. He drinks glasses of wine and smokes a number of cigars. After awhile, Mrs. Pontellier feels like she is being awakened from a dream and realizes that she is quite fatigued. It is almost dawn. Finally getting up from the hammock, Mrs. Pontellier asks her husband if he's going to join her. He replies that he will, after he finishes his cigar.

NovelGuide summary: Mr. Pontellier is surprised to find Edna still outside when he returns from escorting Madame Lebrun home. In a small but no doubt significant exchange-considering the events of the evening, and the novel's title-her distant and unperceiving husband asks her, "Are you asleep?" Edna, with eyes "bright and intense," definitively replies, "No." Although he asks her to come in to the house with him, she refuses, and remains outside, exercising her own will. As if trying to outlast his wife, Mr. Pontellier smokes cigar after cigar next to her. Gradually, Edna succumbs to her need for sleep. She feels "like one who awakens gradually out of a . . . delicious, grotesque, impossible dream . . . ." As described in Chapter VII, then, Edna is once again undergoing what might be called a "negative" "awakening"-an "awakening" to the realities of her present life-as opposed to the "positive" awakening to new possibilities and her own self-direction, to which the nighttime swim began to expose her. As if to underscore her failure to "awaken" to herself, the chapter ends with a scene of tables being turned: as Edna goes in, she asks her husband if he will be joining her. He says he will, as soon as he has finished his last cigar. While the narrator does not record Mr. Pontellier's tone of voice, the comments seem almost scornful, mockingly echoing Edna's earlier self-assertion.

Figure 5: Two reference summaries.

Constituent R-wtd: |I thought I should find you in bed , " ||said her husband , |when he discovered her |lying there . |He had walked up with Madame Lebrun and left her at the house . ||She heard him moving about the room ; |every sound indicating impatience and irritation . |" This is more than folly , " |he blurted out . ' I ca n't |permit you to stay out there all night .||But she could not realize |why or how she should have yielded , feeling as she then did . |He smoked two cigars ; |then he went inside and drank another glass of wine . She tottered up the steps , |clutching feebly at the post before passing into the house . |she asked , |turning her face toward her husband . |

Sentence R-wtd: | I thought I should find you in bed , " said her husband , when he discovered her lying there . | | He had walked up with Madame Lebrun and left her at the house . | | His wife did not reply . | | " This is more than folly , " he blurted out . | | You must come in the house instantly . " | | Edna began to feel like one who awakens gradually out of a dream , a delicious , grotesque , impossible dream , to feel again the realities pressing into her soul . | | She tottered up the steps , clutching feebly at the post before passing into the house . | | she asked , turning her face toward her husband . | | " Just as soon as I have finished my cigar . " |

Figure 6: Two generated summaries. Extracted segments are highlighted in teal, and delineated with |. Constituents are presented with context, whereas sentences extract all text.

Algorithm 1: CONSTITUENTSEGMENTS

Input: sentence parse tree `PT`

1 `const_subtrees :=` [ ]   `// Store constituent subtrees here`

2 `PT, punct_idxs :=` REMOVEPUNCT(`PT`)

3 **foreach** *subtree* `ST` *in* `PT` **do**   `// Find all constituent subtrees.`

4   **if** {`NP, VP`} *in* `ST.children` **then**

5   `STAG = ST`

        `/* Ascend as far as possible in tree, before root S tag          */`

6   **while** `STAG.parent` *in* {`SBAR, S, VP`} **and** `STAG.parent != PT.root` **do**

7   `STAG := STAG.parent`

8   **if** `STAG` *in* {`S, SBAR`} **and** *not* `STAG.children.intersection(`{`VP, NP`}`)`
            then

9   **break**

            `/* If STAG is a VP, no need to break                          */`

10   `const_subtrees := const_subtrees +` [`STAG`]

11   **else if** ISRELATIVECLAUSE(`ST`) **then**

12   `const_subtrees := const_subtrees +` [`STAG`]

  `/* Create words list for each constituent subtree. Avoid duplicating words by`
    `removing subtrees that we add from the original parse tree.                   */`

13 **foreach** *subtree* `ST` *in* `const_subtrees` **do**

14   `WORDS :=` [ ]   `// constituent word lists`

15   **foreach** *left in* `ST.left_siblings` **do**   `// Break up clauses of conjunctions.`

16   **if** `left = CC` **then**

17   `WORDS := WORDS +` [`left.words`]

18   REMOVESUBTREE(`PT, left`)

19   `WORDS := WORDS +` [`ST.words`]

20   REMOVESUBTREE(`PT, ST`)

21 **if** `PT.words` **then**   `// Add any remaining words to another segment`

22   `WORDS := WORDS +` [`PT.words`]

23 `WORDS :=` SPLITNONCONTIGUOUS(`WORDS`)

24 `WORDS :=` SORTBYINDEX(`WORDS`)

25 `WORDS :=` INSERTPUNCTUATION(`WORDS, punct_idxs`)

26 `WORDS :=` CONCATENATESHORTSEGMENTS(`WORDS`)

27 `constituents :=` JOINWORDLISTS(`WORDS`)

Output: constituents $c_1, ..., c_n$

Text:

`` Has n't Miss Gostrey , " she asked , `` said a good word for me ? "

Strether almost wondered -- at such a pace was he going -- if some divination of the influence of either party were what determined Madame de Vionnet 's abstention .

`` Well , I 'm sorry for them . "

all he got in return for it was a sharp `` Better early than late ! "

It 's too late .

The virtuous attachment would be all there before him ; the virtuous attachment would be in the very act of appeal for his blessing ; Jeanne de Vionnet , this charming creature , would be exquisitely , intensely now -- the object of it .

It made Strether guess the more quickly what it might be about which Madame de Vionnet was so urgent .

It was quite as if for an instant the two men looked at each other on this question .

1. Strether is nervous when Chad brings him to see Madame de Vionnet .

○ Present
○ Partially Present
○ Not Present

2. Madame de Vionnet asks Strether whether Miss Gostrey has put in a good word for her .

○ Present
○ Partially Present
○ Not Present

Figure 7: An example HIT showing a segmented oracle summary, and two questions. Reading the summary, we see that we should answer "Present" for both questions. There can be up to 12 questions – we omit here for brevity. Note that in our evaluation, we counted both "Present" and "Partially Present" as a match.