[Data Science](#)[Natural Language Processing](#)[NLP Papers Summary](#)

Day 146: NLP Papers Summary – Exploring Content Selection In Summarization Of Novel Chapters

By Ryan

25th May 2020

No Comments

Objective and Contribution

Proposed a new summarisation task of summarising novel chapters from online study guides. This is much more challenging than the news summarisation due to the length of the source document and the higher level of paraphrasing. The contributions of the paper is as follow:

1. Proposed a new summarisation task of summarising novel chapters



2. Proposed a new metric for aligning sentences in reference summary with sentences in chapter to create good quality “ground-truth” extractive summaries to train our extractive summarisation model. This has proven to improved over previous methods through ROUGE scores and pyramid analysis

Dataset

We collected chapter/summary pairs from five different study guides:

- 1. BarronsBookNotes (BB)
- 2. BookWolf (BW)
- 3. CliffsNotes (CN)
- 4. GradeSaver (GS)
- 5. NovelGuide (NG)

We performed two rounds of filtering to process the data. Firstly, we remove any reference texts with more than 700 sentences as they are too large. Secondly, we remove summaries that are too wordy (compression ratio of less than 2). Our total final chapter / summary pairs is 8088 (6288 / 938 / 862). The training data statistics are shown below. Chapter text, on average, are 7x longer than news articles and chapter summaries are 8x longer than news summaries. In addition, for novel, the average word overlap between summary and chapter is 33.7% whereas for CNN/DailyMail news, it is 68.7%, showcasing a high level of paraphrasing within chapter summaries. This heavy paraphrasing is shown in the example reference summary below.

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:** Word count statistics with total number for summaries and chapter text.



Alignment Experiments

SIMILARITY METRICS

Since the ground-truth summaries are abstractive, we would need to create gold extractive summaries to train our extractive summarisation model. This requires us to align the sentences in the chapter and summary. To align sentences, we first need a metric to measure similarity. Previous work heavily uses ROUGE scores as a similarity metric. However, ROUGE scores assign equal weightings to each word, however, we believe that we should assign higher weight for important words. To incorporate this, we use a smooth inverse frequency weighting scheme and apply this to take the average of ROUGE-1, 2, and L, to generate extracts (R-wtd). We compared this R-wtd approach with other similarity metrics such as ROUGE-1, ROUGE-L, BERT, and unweighted and weighted ROUGE + METEOR (RM). We conducted both automatic evaluation of these similarity metrics using ROUGE-L F1 score and human evaluation. The human evaluation is required to evaluate each reference summary against the aligned sentences. The results are showcase below and R-wtd scored the highest amongst the similarity metrics.

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.

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, '.').



ALIGNMENT METHODS

Once we have established our similarity metric, we now explore the different alignment methods to finally generate our gold extractive summaries. There are two main methods from previous work:

1. *Summary-level alignment*. Selecting the best sentence, comparing to the summary
2. *Sentence-level alignment*. Selecting the best sentence, comparing to each sentence in the summary

For summary-level alignment, we have two variations: selecting sentences until word limit (WL) and selecting sentences until ROUGE score no longer increases (WS summary). For sentence-level alignment, we have two variations: the Gale-Shapley stable matching algorithm and greedy algorithm. The results are displayed below and showcase that the sentence-level stable algorithm performed significantly better than other alignment methods.

Method	RM	R-wtd	RM-wtd	R-1	R-L	BEU
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.


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.

Experiments and Results

For evaluation, we have three extractive models:

1. Hierarchical CNN-LSTM (CB)
2. Seq2seq with attention (K)
3. RNN (N)

We experiment with alignment methods applied at both the word and constituent  since our data analysis shows that summary sentences are often selected from different chapters.

Our evaluation metrics is ROUGE-1, 2, L, and METEOR. Each chapter has 2 – 5 reference summaries and we evaluate our generated summaries against all of them.

RESULTS

The results above compared the performance of three different extractive models as well as the performance difference of using different alignment methods. We can see that our proposed alignment method outperformed the baseline method in all three extractive models. All three models seem to perform similarly using our extractive targets, suggesting the importance of selecting the appropriate method to generate extractive targets. Given the unreliability of ROUGE, we perform human evaluation and compute the pyramid score of each alignment methods on our best performing model (CB). The crowd workers are asked to identify which generated summary best convey the sampled reference summary content. The results are displayed below.

Conclusion and Future Work

We have shown that sentence-level, stable-matched alignment method with R-wtd similarity metric performed better than previous method of computing gold extractive summaries. However, there seem to be a contradictory in automatic and human evaluation on whether extraction is better at the sentence or constituent level. We speculate that this might be because we didn't include the additional context when scoring the summaries of extracted

constituents and so the irrelevant context didn't go against the system whereas in the h ^ an evaluation, we do include sentence context and so fewer constituents are included in the generated summary.

In future work, we plan on examining how we can combine constituents to make fluent sentences without including irrelevant context. We would also like to explore abstractive summarisation, to examine if language models would be effective in our domain. This could be challenging as language models typically has a limit of 512 tokens. The truncation of our documents might hurt the performance of our novel chapter summarisation model.

Source: <https://arxiv.org/pdf/2005.01840.pdf>

Ryan

Data Scientist

Previous Post

< Day 145: NLP
Papers Summary -
SUPERT: Towards
New Frontiers in
Unsupervised
Evaluation Metrics
for Multi-
Document
Summarization

Next Post

Day 147: NLP
Papers Summary -
Two Birds, One
Stone: A Simple,
Unified Model for
Text Generation
from Structured
and Unstructured
Data





[Data Science](#) [Natural Language Processing](#) [NLP Papers Summary](#)

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan

30th December 2020





[Data Science](#) [Implementation](#) [Natural Language Processing](#)

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan

29th December 2020



