Gold Corpus for Telegraphic Summarization

Chanakya Malireddy
LTRC, IIIT Hyderabad
chanakya.malireddy
@research.iiit.ac.in

Srivenkata N M Somisetty
IIIT Hyderabad
mounikas@gmail.com

Manish Shrivastava
LTRC, IIIT Hyderabad
m.shrivastava@iiit.ac.in

Abstract

Most extractive summarization techniques operate by ranking all the source sentences and then select the top-ranked sentences as the summary. Such methods are known to produce good summaries, especially when applied to news articles and scientific texts. However, they do not fare so well when applied to texts such as fictional narratives, which do not have a single central or recurrent theme. This is because usually the information or plot of the story is spread across several sentences. In this paper, we discuss a different summarization technique called Telegraphic Summarization. Here, we do not select whole sentences, rather pick short segments of text spread across sentences, as the summary. We have tailored a set of guidelines to create such summaries and, using the same, annotate a gold corpus of 200 English short stories.

1 Introduction

The purpose of summarization is to capture all the useful information from the source text in as few words as possible. Extractive summarization involves identifying parts of the text which are important. Such summaries are usually generated by ranking all the source sentences, according to some heuristic or metric, and then selecting the top sentences as the summary. Most extractive summarization systems have been developed for domains such as newswire articles (Lee et al., 2005), encyclopedic and scientific texts (Teufel and Moens, 2002). They work well in such domains because these texts revolve around a central theme and the information is often enforced by reiteration across several sentences. However, fictional narratives do not talk about a single topic. They describe a sequence of events and often contain dialogue. Information is not repeated and each sentence contributes to developing the plot further. Hence, selecting a subset of such sentences does not accurately capture the story.

In this paper, we focus on telegraphic summarization. Telegraphic summary does not contain whole sentences, instead, shorter segments are selected across sentences, and reads like a telegram. We have described a set of guidelines to create such summaries. These guidelines were used to annotate a gold corpus of 200 English short stories. The dataset can be very useful to gain an insight into story structures. No such corpus already exists, to the best of our knowledge. Formally, the paper makes the following contributions:

1. Discusses drawbacks of applying traditional extractive summarization methods to narrative texts.
2. Describes a set of guidelines to generate telegraphic summaries.
3. Provides a gold corpus of 200 short stories and their telegraphic summaries annotated using these guidelines.
4. Provides abstractive summaries for 50 stories and a set of 45 multiple-choice questions (MCQs) for evaluation purposes.

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 71–77
Santa Fe, New Mexico, USA, August 20, 2018.

The paper is organized as follows. Section 2 discusses existing and related work. Section 3 describes the data collection and summarization process. Section 4 discusses the analysis performed on the dataset. Section 5 discusses conclusions and future work.

2 Related Work

Automatic text summarization was first attempted in the middle of the 20th century (Luhn, 1958). Since then it has been applied to several domains and corpora, such as news articles (Lee et al., 2005), scientific articles (Teufel and Moens, 2002), and blogs (Hu et al., 2007).

News articles have been the focus of summarization systems for a long time because of the vast practical applications. In fact, most datasets available today are built from news corpora. However, a comparative study has shown that a single summarization technique does not perform equally well across all domains (Ceylan et al., 2010). Therefore, separate systems have to be built to deal with the domain of fiction and its nuances, and news corpora based datasets are not sufficient to train and evaluate the same.

There has been research on short fiction summarization (Kazantseva and Szpakowicz, 2010), fairy tales (Lloret and Palomar, 2009) and whole books (Mihalcea and Ceylan, 2007). But the aforementioned work in short fiction summarization had a different objective - helping a reader decide whether one would be interested in reading the complete story. Hence it contains just enough information to help the reader decide but does not reveal the entire plot. However, our dataset aims to summarize the entire plot. This is useful to learn plot structures and story organization.

Turney (2000) and the KEA algorithm by Witten et al. (1999) have attacked the problem of key-phrase extraction. But the key-phrases extracted by them do not form a cohesive summary and just try to list the major themes discussed in the article and therefore cannot be applied to the domain of fiction.

Grefenstette (1998) proposed the use of sentence shortening to generate telegraphic texts that would help a blind reader skim a page (using text-to-speech). He provided eight levels of telegraphic reduction. The first (and the most drastic) generated a stream of all the proper nouns in the text. The second generated all nouns present in the subject or object position. The third, in addition, included the head verbs. The least drastic reduction generated all subjects, head verbs, objects, subclauses, prepositions and dependent noun heads. Since then Jing (2000), Riezler et al. (2003) and Knight and Marcu (2000) have explored statistical models for sentence shortening that, in addition, aim at ensuring grammaticality of the shortened sentences. Intuitively it appears that sentence-shortening can allow more important information to be included in a summary. However, Lin (2003) showed that statistical sentence-shortening approaches like Knight and Marcu (2000) resulted in significantly worse content selection. He concluded that pure syntax-based compression does not improve overall summarizer performance, even though it performs well at the sentence level. Which is why there is a need for semantically aware techniques. Our dataset provides the tools to evaluate and, in the future, maybe even train such algorithms.

3 Data Construction

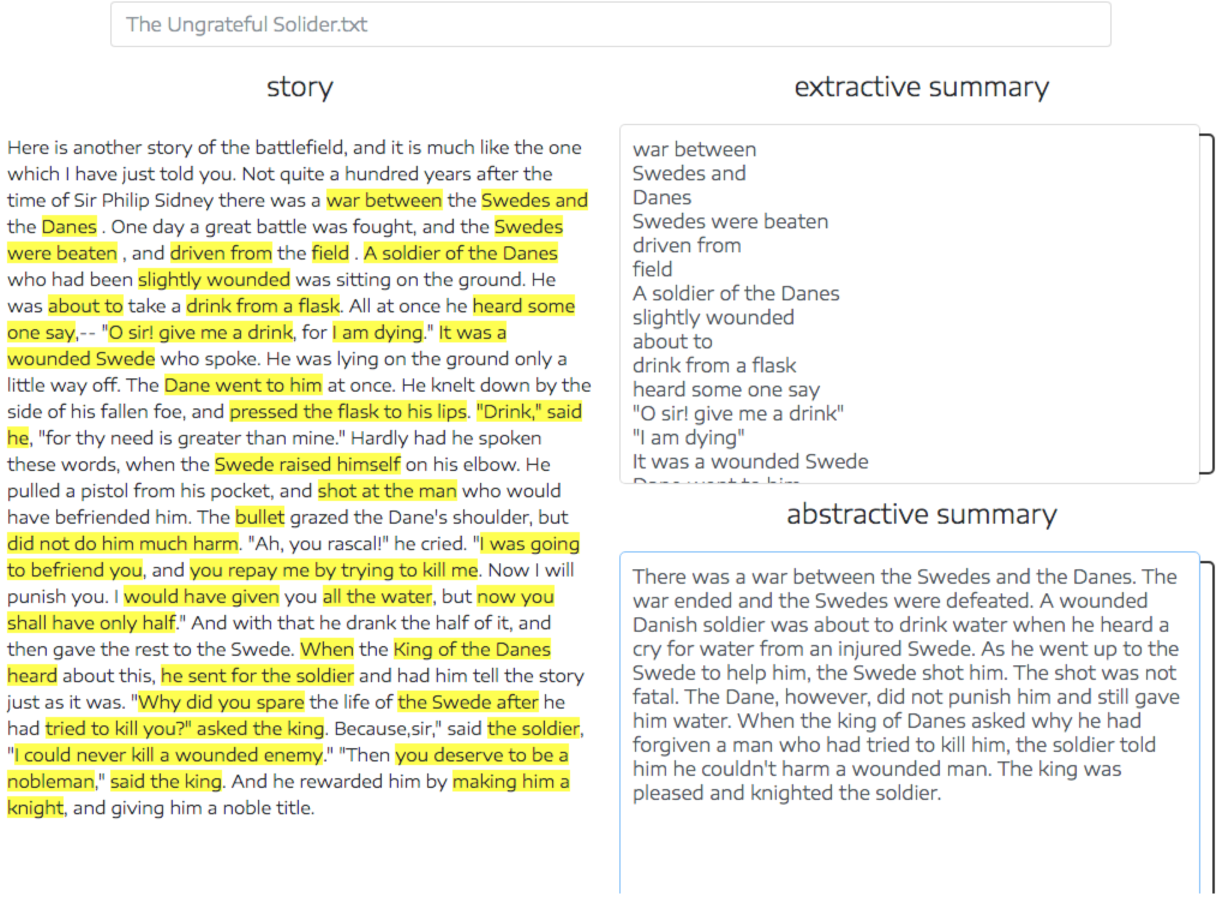
3.1 Collection and Preprocessing

We collected English short stories containing 300 to 1100 words, available in the public domain[1]. 200 stories were then randomly picked, after ensuring that works from at least 20 different authors had been selected to keep the dataset diverse in terms of genre and writing style. As a result, the dataset spans 39 authors. These stories were then manually processed to remove any spelling, grammatical and encoding errors that might have crept in.

3.2 Summarization

The summarization was performed by 5 annotators. The annotators are not native English speakers but are fluent in the language. They summarized 40 stories and cross-annotated 4 stories each (1 from each remaining annotator), to help calculate inter-annotator agreement. Each annotator also performed

[1] https://americanliterature.com/short-story-library



The Ungrateful Solider.txt

story

Here is another story of the battlefield, and it is much like the one which I have just told you. Not quite a hundred years after the time of Sir Philip Sidney there was a war between the Swedes and the Danes . One day a great battle was fought, and the Swedes were beaten , and driven from the field . A soldier of the Danes who had been slightly wounded was sitting on the ground. He was about to take a drink from a flask. All at once he heard some one say,-- "O sir! give me a drink, for I am dying." It was a wounded Swede who spoke. He was lying on the ground only a little way off. The Dane went to him at once. He knelt down by the side of his fallen foe, and pressed the flask to his lips. "Drink," said he, "for thy need is greater than mine." Hardly had he spoken these words, when the Swede raised himself on his elbow. He pulled a pistol from his pocket, and shot at the man who would have befriended him. The bullet grazed the Dane's shoulder, but did not do him much harm. "Ah, you rascal!" he cried. "I was going to befriend you, and you repay me by trying to kill me. Now I will punish you. I would have given you all the water, but now you shall have only half." And with that he drank the half of it, and then gave the rest to the Swede. When the King of the Danes heard about this, he sent for the soldier and had him tell the story just as it was. "Why did you spare the life of the Swede after he had tried to kill you?" asked the king. Because,sir," said the soldier, "I could never kill a wounded enemy." "Then you deserve to be a nobleman," said the king. And he rewarded him by making him a knight, and giving him a noble title.

extractive summary

war between
Swedes and
Danes
Swedes were beaten
driven from
field
A soldier of the Danes
slightly wounded
about to
drink from a flask
heard some one say
"O sir! give me a drink"
"I am dying"
It was a wounded Swede

abstractive summary

There was a war between the Swedes and the Danes. The war ended and the Swedes were defeated. A wounded Danish soldier was about to drink water when he heard a cry for water from an injured Swede. As he went up to the Swede to help him, the Swede shot him. The shot was not fatal. The Dane, however, did not punish him and still gave him water. When the king of Danes asked why he had forgiven a man who had tried to kill him, the soldier told him he couldn't harm a wounded man. The king was pleased and knighted the soldier.

Figure 1: An example of telegraphic and abstractive summaries created according to the guidelines. The story is displayed in the left panel, the telegraphic summary is highlighted and also listed in the box titled 'extractive summary' and the abstractive summary is shown below it.

abstractive summarization on 10 stories. Unlike extractive summarization, abstractive summaries need not contain the same words as used in the source and are instead written by the annotator in their "own words" based on their understanding of the text. Abstractive summary for a story was not provided by the same annotator who generated its telegraphic summary. An example is shown in Figure 1.

Guidelines followed for telegraphic summarization:

1. A segment is defined as a continuous span of words in the source, chosen as a part of the summary.
2. A word should not be fragmented e.g., if the word "breaking" appears in the source, it should not be broken into fragments like "break".
3. Each segment should be relevant to the plot, try to advance the story and have some continuity with the preceding and the following segment.
4. Segments extracted from dialogues or parentheses should be enclosed in quotes or parentheses respectively.
5. Segments should be arranged in the same order as they appear in the story.
6. The summary should be minimal. If multiple segments mean the same thing, pick the shortest. Adjectives, adverbs, and modifiers are not to be included if they are not relevant to the plot.
7. When the segments are read in sequence the plot should be apparent and unambiguous.

Guidelines followed for abstractive summarization:

1. Summaries should be written from a third party perspective e.g., “This story is about a girl...”
2. Summaries should only discuss the plot and try to avoid inferences and opinions not immediately apparent from the story.
3. Summaries should maintain the same order of events as they occur in the source text.

4 Data Analysis and Evaluation

The length of the stories varies from 300 to 1100 words, with the average length being 650. The average summarization factor (length of summary/length of the story) is 0.37 and 0.36 for telegraphic and abstractive summaries respectively.

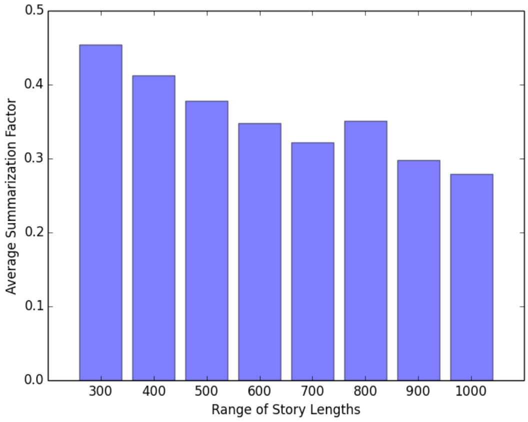


Figure 2: Summarization Factor vs Story Length

From Figure 2 we can see that the summarization factor tends to be high for very short stories since nearly every word is important. It reduces gradually as the length of the story increases because longer stories tend to be more descriptive and contain extraneous information not relevant to the central plot.

4.1 Quantitative Evaluation

We used the Alpha metric proposed in Krippendorff (1980) to measure the inter-annotator agreement. Alpha was computed based on 20 cross-annotated stories and found to be 0.73.

We generated 50 summaries using popular online extractive summarization tools, Smmry[2] and Resoomer[3], which generate summaries by ranking and selecting the top sentences (after the selection step the sentences are re-arranged according to the source order). These summaries have the same summarization factor as the corresponding telegraphic summaries.

$$Rouge_N = \frac{\sum_{s \epsilon \{ReferenceSummaries\}} \sum_{gram_n \epsilon S} Count_{\text{match}}\left(gram_n\right)}{\sum_{s \epsilon \{ReferenceSummaries\}} \sum_{gram_n \epsilon S} Count\left(gram_n\right)} \tag{1}$$

ROUGE-N score (Lin, 2004) is a popular metric used to evaluate summaries produced by a system. ROUGE-N recall is computed as shown in Eq. 1, where N stands for the length of the n-gram, $gram_n$ and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. ROGUE-N precision can be calculated by replacing the denominator in Eq. 1 by the total number of n-grams present in all the reference summaries instead of system summaries. The F1 score is defined as the Harmonic Mean of precision and recall.

[2] https://smmry.com/

[3] https://resoomer.com/en/

In our case, there was only one reference (the abstractive summary) and one summary from each system - Telegraphic, Smmry, and Resoomer. We report the average F1 score, after removing stop words and stemming, for N = 1,2,3,4 in Table 1.

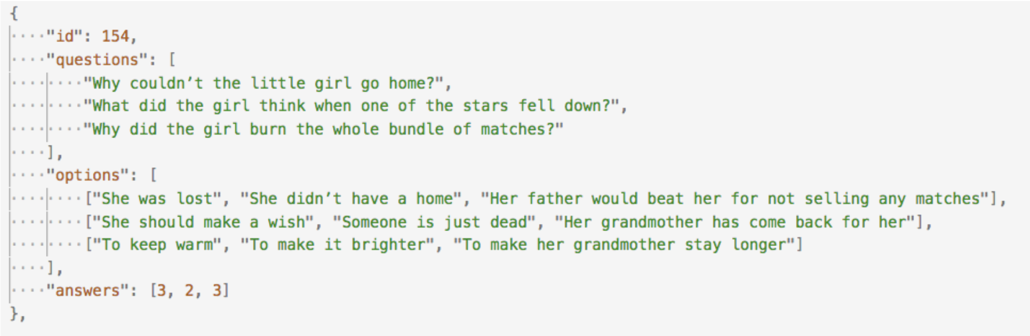
	N=1	N=2	N=3	N=4
Telegraphic	0.582	0.258	0.108	0.051
Smmry	0.498	0.184	0.092	0.056
Resoomer	0.483	0.179	0.093	0.057

Table 1: ROUGE-N F1 score

The higher F1 score for N = 1,2,3 telegraphic summaries indicates that they are more adequate in terms of capturing relevant content. Since telegraphic summaries select shorter segments of text, they can retain more information while maintaining the same summarization factor as their sentence-level counterparts. The ROUGE-4 score for Smmry and Resoomer summaries is slightly higher because they select entire source sentences and are therefore likely to have more 4-gram overlaps.

4.2 Qualitative Evaluation

High-order n-gram ROUGE measures try to judge fluency to some degree but since ROUGE is based only on the content overlap, it can only measure adequacy and not coherence. In order to gain an insight into the coherence of the summaries, we made a set of 45 MCQs from 15 stories in the dataset. Questions were not set by the same annotator who generated the corresponding telegraphic summary. The test was administered to three groups of two participants each. Each group was allotted the summaries produced by a single system - Telegraphic, Smmry or Resoomer.



```
{
    "id": 154,
    "questions": [
        "Why couldn’t the little girl go home?",
        "What did the girl think when one of the stars fell down?",
        "Why did the girl burn the whole bundle of matches?"
    ],
    "options": [
        ["She was lost", "She didn’t have a home", "Her father would beat her for not selling any matches"],
        ["She should make a wish", "Someone is just dead", "Her grandmother has come back for her"],
        ["To keep warm", "To make it brighter", "To make her grandmother stay longer"]
    ],
    "answers": [3, 2, 3]
},
```

Figure 3: An example set of questions for the story ‘The Little Match Girl’.

An example set of questions for the story ‘The Little Match Girl’ is shown in Figure 3. The ‘id’ field refers to the unique id we assign to each story in the dataset. For each story, we made a set of 3 multiple-choice questions with 3 options each. The answers refer to the index of the correct option in the list. Apart from the given options, the participants were allowed to choose option 4, “Can’t say”, if they could not answer a question based on the summary. Average scores are reported in Table 2.

	Correct	Incorrect	Can’t say
Telegraphic	88.9%	2.2%	8.9%
Smmry	62.2%	4.4%	33.4%
Resoomer	60.0%	2.2%	37.8%

Table 2: Questionnaire Results

Higher scores on the questionnaire indicate that the telegraphic summaries were more coherent and allowed the reader to understand the story better. Participants who read the Smmry and Resoomer summaries were unable to understand the story and chose "Can't say" as the answer for nearly a third of the questions.

5 Conclusion and Future Work

In this paper, we highlight the shortcomings of applying traditional extractive summarization techniques to narrative texts and show how telegraphic summarization can be used to overcome these shortcomings.

We defined a set of guidelines to help generate telegraphic summaries. Using the same, we then construct a corpus of 200 English short stories and their telegraphic summaries. 50 abstractive summaries and 45 MCQs are also provided for evaluation purposes. This corpus has been made public [4] and can be used as a gold standard to evaluate such summarization tasks. The MCQs can also be used to evaluate QA systems.

In future, we intend to extend this corpus by adding more stories. We plan on developing algorithms to automatically generate high-quality telegraphic summaries and an extended corpus could be used as training data for supervised techniques.

References

Hakan Ceylan, Rada Mihalcea, Umut Özertem, Elena Lloret, and Manuel Palomar. 2010. Quantifying the limits and success of extractive summarization systems across domains. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 903–911. Association for Computational Linguistics.

Gregory Grefenstette. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Working notes of the AAAI Spring Symposium on Intelligent Text summarization*, pages 111–118. The AAAI Press Menlo Park, CA.

Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2007. Comments-oriented blog summarization by sentence extraction. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 901–904, New York, NY, USA. ACM.

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 310–315, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anna Kazantseva and Stan Szpakowicz. 2010. Summarizing short stories. *Comput. Linguist.*, 36(1):71–109.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. *AAAI/ IAAI*, 2000:703–710.

Klaus Krippendorff. 1980. *Content analysis: an introduction to its methodology*. Sage commtext series. Sage Publications.

Chang-Shing Lee, Zhi-Wei Jian, and Lin-Kai Huang. 2005. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(5):859–880.

Chin-Yew Lin. 2003. Improving summarization performance by sentence compression: a pilot study. In *Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11*, pages 1–8. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Elena Lloret and Manuel Palomar, 2009. *A Gradual Combination of Features for Building Automatic Summarisation Systems*, pages 16–23. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

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.

[4] https://github.com/m-chanakya/shortstories

Stefan Riezler, Tracy H King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 118–125. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Peter D Turney. 2000. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. KEA: practical automatic keyphrase extraction. *CoRR*, cs.DL/9902007.