# Understanding the Behaviour of Neural Abstractive Summarizers using Contrastive Examples

**Krtin Kumar**
School of Computer Science
McGill University
krtin.kumar@mail.mcgill.ca

**Jackie Chi Kit Cheung**
School of Computer Science
McGill University
jcheung@cs.mcgill.ca

## Abstract

Neural abstractive summarizers generate summary texts using a language model conditioned on the input source text, and have recently achieved high ROUGE scores on benchmark summarization datasets. We investigate how they achieve this performance with respect to human-written gold-standard abstracts, and whether the systems are able to understand deeper syntactic and semantic structures. We generate a set of contrastive summaries which are perturbed, deficient versions of human-written summaries, and test whether existing neural summarizers score them more highly than the human-written summaries. We analyze their performance on different datasets and find that these systems fail to understand the source text, in a majority of the cases.

## 1 Introduction

Open-domain abstractive summarization is a longstanding goal of the field of automatic summarization. Compared to extractive techniques, abstraction offers the potential to generate much more useful summaries by simplifying and rephrasing the source text (Knight and Marcu, 2002), and furthermore by aggregating information and performing operations which are not possible with extractive techniques (Genest and Lapalme, 2012; Carenini and Cheung, 2008).

Recently, a number of abstractive summarization systems based on neural sequence-to-sequence architectures have been proposed (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Paulus et al., 2018; Chen and Bansal, 2018). These systems learn a compressed representation of the source text using an encoder, then generate the output summary using a conditional decoder. Such neural abstractive systems have achieved very good ROUGE scores on different datasets.

| **Source** |
|---|
| A former Iraqi army chief of staff being investigated in denmark for war crimes is believed to be back in **Iraq**, one of his sons said Tuesday. |
| **Contrastive 1** |
| Iraqi general missing **in** denmark believed to be back **from** Iraq. |
| **Contrastive 2** |
| Iraqi general missing from **Iraq** believed to be back in Iraq. |

Table 1: Examples of generated contrastive summaries. Bold indicate switched words

.

Our interest in this paper is to investigate how these abstractive systems achieve such results, and whether they represent progress towards language understanding and generation. ROUGE arguably provides a limited view of the performance of such systems, as they only relate the system summary to a fixed number of gold-standard summaries. We propose a novel method to directly test the abstractive summarizers in terms of how they score potential candidate summaries, viewing them as conditional language models. This allows us to test whether the summarizers favour output summaries with specific desired qualities, such as generating a summary that is semantically consistent and entailed by the source text.

We test how well the neural abstractive summarizers distinguish human-written abstracts from contrastive distractors, which are clearly incorrect summaries that are generated using a rule-based procedure. Table 1 shows contrastive examples which are clearly incorrect[1]. In majority of source texts, we are able to find a contrastive example that scores more highly than the gold-standard summary.

[1] For other NLP tasks, others have proposed a similar notion called *adversarial* examples (Jia and Liang, 2017). Since the term ‘adversarial’ has traditionally implied learning to specifically attack the weakness of a model, which we do not do, we refrain from using the word ‘adversarial’.

*Proceedings of NAACL-HLT 2019*, pages 3949–3954
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

Our work demonstrates the difficulty of controlling expressive neural abstractive systems to produce correct and fluent output. It also underscores the need to revisit fundamental issues in summarization evaluation for neural abstractive models, so that a comprehensive evaluation scheme that captures all relevant aspects of summary quality can be developed. Our code for generating contrastive summaries is available online.[2]

## 2 Related Work

Most work in neural abstractive summarization has focused on optimizing ROUGE, whether implicitly by maximum likelihood training or explicitly by reinforcement learning. While this could certainly capture aspects of the content selection problem, we believe that the focus should now shift towards semantic correctness and readability.

Cao et al. (2018) took a step in this direction through their fact-aware neural abstractive summarization system. They use fact descriptions of the source as additional features for the summarizer, and showed improved faithfulness according to human judgments. Multi-task learning is another approach used by Pasunuru et al. (2017) to reduce semantic errors in the generated summaries. They jointly learn summarization and entailment generation tasks, using different encoders but a shared decoder.

A number of automatic evaluation metrics have shown high correlation with human judges (Liu and Liu, 2008; Graham, 2015), but these results are either restricted to extractive systems or were performed with respect to human-generated summaries. Correlation values are significantly reduced when performed on abstractive summarization systems and datasets (Toutanova et al., 2016).

## 3 Generating Contrastive Summaries

In this section, we describe our method for evaluating summarization systems based on whether they can separate human-written gold summaries from automatically generated contrastive summaries. We define a contrastive summary to be similar to a gold summary, except it contains a perturbation. The perturbation results in either a semantic discrepancy, where facts in source and summary do not corroborate, or a readability issue, where issues with grammar or fluency renders the summary clearly incorrect. Below, we first describe our basic method of introducing these discrepancies, then describe a number of restrictions we apply to ensure that the generated contrastive summaries are of high quality.

| **Rule** | **Switching Criteria** |
|---|---|
| Noun | NN, NNP or NNS must match. For NNP child DET if present is also switched. |
| Preposition | For IN tags: parents and their dependencies must match. |
| Verb | VBP, VBG, VBZ, VBN, VBD or VB must match |
| Adjective | JJ, JJR or JJS must match |

Table 2: Rules for selecting words to switch when generating the contrastive summaries. The tags are as per the Penn Tree-bank (Santorini, 1990).

**Perturbation by switching words.** Given pairs of source texts and gold summaries, we generate multiple contrastive summaries for each source text by perturbing its associated gold summary. There are many types of possible perturbations, but we focus on two strategies: 1) switching words within a gold summary, and 2) replacing a word in gold summary by a word from the source text. We chose these types of perturbations as they are likely to result in "difficult" contrastive summaries that contain words which are likely to appear in a reasonable summary of the source text, but which are nevertheless incorrect.

In order to select the words to be swapped, we apply four rules, separated by syntactic category, as shown in Table 2, using the dependency parse of the texts (Manning et al., 2014). We switch words either within a gold summary, or from the source text and use a single rule at a time for generating a contrastive summary. For example, if the POS tag *NNS* is matched between a word *'sides'* in the source text and the word *'combatants'* in the gold summary, then the Noun rule would apply, and the words would be switched to obtain the contrastive summary.

Further, for the Noun and Verb categories, the switched words may not match in number or verb conjugation. We use SimpleNLG (Gatt and Reiter, 2009) to convert the word to the appropriate inflectional form of the destination's POS tag.

[2] https://github.com/krtin/ContrastiveSummaries

**Further restrictions.** We place a number of restrictions on the words switched, to ensure that the generated summary is contrastive compared to the gold summary. We do not allow switching of the same words. We also do not allow words to be switched if they are separated by any of the following: *'or'*, *'and'* or *','*, as these are likely to be commutative operators.

Furthermore, we only allow switching of words from the source text if the context of the words to be switched sufficiently differ from each other. We compute unigram overlap around a context window of size 2 on each side, and allow a switch when the overlap proportion is less than 0.65. These settings were determined by manual inspection of the generated summaries, and allowed us to reduce the number of examples where generated summary is not contrastive, without significantly reducing the number of generated summaries.

We will describe the results of a human verification study in Section 5, in which we ask human raters to check the quality of our contrastive summaries.

## 4 Evaluation

We apply our set of contrastive summaries to evaluate a number of neural abstractive summarizers. For each summarizer under evaluation, we assume access to a conditional language model which defines a probability distribution over words conditioned on the source. Formally, such a language model is given by Equation 1:

$$P(y_i|\theta, \mathcal{S}, y_1...y_{i-1}), \qquad (1)$$

here $\mathcal{S}$ is the source, $\theta$ are model parameters, $y_i \in V_{sm}$ represents the $i^{th}$ word in the summary, $V_{sm}$ is the vocabulary space of the summary and $P$ is the conditional probability. $\mathcal{S} \subseteq (s_1, ..., s_n)$ where $s_i \in V_{so}$, $V_{so}$ is the vocabulary space of the source and $n$ is the maximum source length. Further, we use Equation 2 as our scoring function,

$$p(y) = \sum_{i=1}^{m} \log P(y_i|\theta, \mathcal{S}, y_1...y_{i-1}), \qquad (2)$$

here $m$ is maximum summary length and $y$ represents a gold or contrastive summary. For a given triple of source, gold (g) and contrastive (c) summary, if $p(g) > p(c)$, then we label the triple *'dodged'*, since the summarizer successfully dodged the generated contrastive summary. If a system is able to dodge all contrastive summaries generated from a source and gold summary tuple, then we label the tuple as *'escaped'*.

## 5 Experiments

**Datasets.** We experimented on two datasets, for two abstractive summarization tasks. The first is a short summarization task, where the summary is one sentence long, for which we use the **Gigaword** corpus (Graff et al., 2011; Napoles et al., 2012). We use the scripts provided by Rush et al. (2015) to process the Gigaword corpus, which contains the first sentence of the article and the headline as source and gold summary pairs. The test set contains about 250K source-summary pairs from which we randomly selected 10K pairs and generated 509K contrastive summaries.

The second is a long summarization task, in which the summary consists of multiple sentences. We use the **CNN/Dailymail** corpus, where the highlights of the articles are used as the gold summary (Hermann et al., 2015). We used the scripts from Nallapati et al. (2016) to get the data and use the non-anonymized version like See et al. (2017). We use 11.49K test pairs and were able to generate 563K contrastive summaries.

**Models.** We analyze and evaluate three state-of-the-art neural abstractive summarization systems: ABS+ (Rush et al., 2015), GTP (See et al., 2017) and FAS (Chen and Bansal, 2018). The **ABS+** system uses an attention-based neural language model as an encoder and a feed-forward neural network for decoding, and is trained on the Gigaword corpus. The **GTP** system is a seq2seq model with attention on the encoder and a pointer-generator mechanism to choose words from the source in the decoder and, is trained on the CNN/Dailymail corpus. **FAS** uses reinforcement learning algorithm to extract the most important sentences from the source text, and then summarizes each sentence using a similar architecture as GTP on the CNN/Dailymail corpus.

These systems have performed well on small and large text summarization tasks, and have open-source implementation available from the authors. We would have liked to test other relevant systems (Pasunuru et al., 2017; Cao et al., 2018), but were unable to obtain their implementations.

**Experimental Details.** The CNN/Dailymail corpus has a large source length, thus the set of

|  | Rule | Dodged (%) |  |  |
|---|---|---|---|---|
|  |  | CNN |  | Gigaword |
|  |  | GTP | FAS | ABS+ |
| Gold | Noun | 98.8 | 96.5 | 49.5 |
|  | Prep | 97.8 | 96.9 | 55.4 |
|  | Verb | 99.0 | 98.5 | 48.7 |
|  | Adj | 97.8 | 95.6 | 55.7 |
| Source | Noun | 94.3 | 88.5 | 47.9 |
|  | Prep | 92.0 | 88.2 | 50.1 |
|  | Verb | 94.9 | 91.9 | 50.2 |
|  | Adj | 92.6 | 87.2 | 49.7 |

Table 3: Rule-wise Performance, here Source and Gold are based upon rule perturbations in Section 3

| Model | Dataset | Dodged | Escaped |
|---|---|---|---|
| GTP | CNN | 96.3% | 29.8% |
| FAS | CNN | 92.9% | 12.2% |
| ABS+ | Gigaword | 48.6% | 10.8% |

Table 4: Performance on *dodged*, *escaped* metrics.

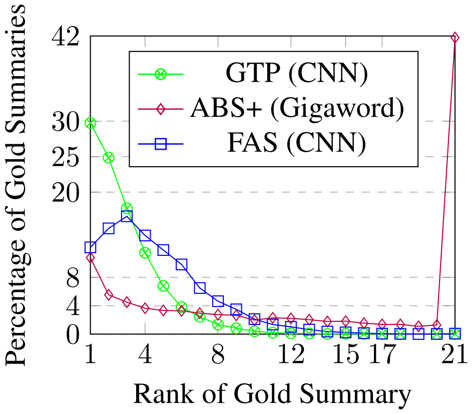

Figure 1: Distribution of Gold Summary Rank

contrastive summaries is very large. To restrict the number of contrastive summaries we randomly select approximately 50 generated summaries, while maintaining the rule-wise distribution. The rule-wise distribution was estimated based upon contrastive summaries, generated from a subset of 100 gold standard summaries from CNN/Dailymail corpus.

In order to correctly evaluate the FAS system, for each extracted sentence we generate all sentences in the gold summary, and pick the set of summaries with the maximum probability.

**Human verification.** To ensure that we are generating incorrect contrastive summaries, 200 randomly sampled summaries from the Gigaword corpus were evaluated by a human annotator, to verify if a semantic discrepancy or a readability issue was present. We ensured that we sample equally across all the 8 rules, and that we restrict our set of contrastive summaries which the ABS+ system was not able to *'dodge'*. We found that 49.5% had a readability issue while 43.5% had a discrepancy issue, and 93% of the examples had at least one of these issues. This indicates that the vast majority of our contrastive examples are "true negatives"; i.e., a perfect summarization system should score them lower than the gold standard summary.

## 6 Results

We summarize our results in Table 4, and report rule-wise results in Table 3. We also include examples where the ABS+ system is unsuccessful in dodging the generated contrastive summaries, in Table 5. Since these metrics directly evaluate the posterior distribution of a summarizer, it allows us to explicitly recognize problematic examples for a model. We also look at what percentage of gold summaries lie across different ranks of gold summaries in Figure 1. This gives us an insight into distribution of gold summaries, across different ranks. The rank of a gold summary is 1 plus the number of contrastive summaries that scored higher than it.

**Dodged and Escaped.** On the CNN/Dailymail dataset, we find both the models were able to dodge most of the contrastive summaries, but a large number of summaries had at least a few contrastive summaries which scored higher than the gold summary, as reflected by the *escaped* metric.

The FAS model performs worse than GTP model, this might be because the abstraction model only observes one sentence, and thus the probability of observing a word outside the source sentence is higher for the contrastive summaries.

**Rule-wise Analysis.** The GTP and FAS models perform better on rules which switch words within the gold-standard summary. Thus, the decoder LSTM has captured the data distribution very well for words within the summary but is not generalizing for words outside it. This suggest that using words outside the source vocabulary might help in generating harder contrastive examples. The ABS+ model is better in capturing data distributions of prepositions and adjectives. This points towards biases towards particular distributions and can be helpful in further improving these models.

**Rank of Gold Summary.** As shown in Figure 1, almost all the gold summaries have rank lower than 8 for the GTP model, while a large percentage of gold summaries have rank greater than 20, for the ABS+ model. The maximum rank in both the cases is of the order 500K, which is the size of contrastive summaries (Section 5). We suspect that this might be due to behaviourally extractive nature of GTP model, which allows it to easily distinguish any perturbations in the contrastive summaries.

## 7 Conclusion

We proposed to analyze existing neural abstractive summarizers by testing how they score contrastive summaries, compared to gold-standard ones. For the majority of the gold-standard summaries, we were able to find contrastive examples which score more highly according to current state-of-the-art systems. These examples can be useful not only in evaluating the performance of these systems, but also for improving these systems in the future.

## References

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization.

Giuseppe Carenini and Jackie Chi Kit Cheung. 2008. Extractive vs. nlg-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

Pierre-Etienne Genest and Guy Lapalme. 2012. Fully abstractive approach to guided summarization. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 354–358. Association for Computational Linguistics.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword, linguistic data consortium.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Feifan Liu and Yang Liu. 2008. Correlation between rouge and human evaluation of extractive meeting summaries. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*, pages 201–204. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny 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.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, 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. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Ramakanth Pasunuru, Han Guo, and Mohit Bansal. 2017. Towards improving abstractive summarization via entailment generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 27–32. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. *International Conference on Learning Representations*.

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

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, page 570.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In (See et al., 2017), pages 1073–1083.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.

## A Appendices

| | |
|---|---|
| **Example 1** | |
| **Source** | The Asian economic contagion that contributed overnight to the year's worst loss on wall street returned home with the opening of trading Wednesday, helping to drag some markets to levels they hadn't seen for months, if not years. |
| **Gold** | Asian markets fall after big drop on wall street. |
| **Contrastive** | Asian markets fall after big drop on **street wall**. |
| **Problems** | The system failed to understand that wall street has real-world significance |
| **Example 2** | |
| **Source** | Panama's colon duty free area, latin america's biggest re-export zone, ground to a halt monday as business leaders shut their doors to protest a tax hike. |
| **Gold** | Business leaders launch strike in Panama's free zone by James. |
| **Contrastive** | Business **strike** launch **leaders** in Panama's free zone by James. |
| **Problems** | Failed to understand that business leaders belong to a single entity |
| **Example 3** | |
| **Source** | Iran stood firm on its position towards terrorism and the middle east peace process in talks with an EU mission here … |
| **Gold** | Iran refuses to budge on terrorism, peace process in talks with EU by Laurent. |
| **Contrastive** | Iran refuses to budge on terrorism, peace process **by** talks with EU **in** Laurent. |
| **Problems** | Improper usage and understanding of Prepositions |
| **Example 4** | |
| **Source** | The rear door of a Russian-made cargo plane crammed with Congolese soldiers and their families flew open in midair on Thursday night, 33,000 feet above the jungles of Congo, dropping scores of passengers down a ramp and into the sky, survivors said. |
| **Gold** | Passengers fall from plane over Congo; death toll unclear |
| **Contrastive** | Passengers fall **over** plane **from** Congo; death toll unclear |
| **Problems** | Failed to understand that falling over a plane is improbable in real-world context |
| **Example 5** | |
| **Source** | The leader of the separatist Georgian black sea region of Abkhazia on Monday rejected Tbilisi's insistence that railway **traffic** via the region would only be restored when refugees displaced by war are allowed to return and their safety is ensured. |
| **Gold** | Georgia's breakaway Abkhazia rejects conditions for restoring **rail** traffic. |
| **Contrastive** | Georgia's breakaway Abkhazia rejects conditions for restoring **traffic** traffic. |
| **Problems** | A simple repetition issue, also pointed out by the authors (Rush et al., 2015) |
| **Example 6** | |
| **Source** | Incheon, a port city in the republic of Korea (rok), plans to restore its century-old china town, which was destroyed during the Korean war (1950-60), and to build it into a **tourism** attraction by inviting investment from china … |
| **Gold** | Rok city wants to rebuild china town with Chinese **investment**. |
| **Contrastive** | Rok city wants to rebuild china town with Chinese **tourism**. |
| **Problems** | The model lacks understanding of the source text |

Table 5: Examples of contrastive summaries, that ABS+ system was not able to dodge. Bold indicate switched words