ex-GPT: An Extractive-Abstractive Summarization Framework with a Sentence Embeddings Twist

Lorenzo Minto **Jakub Kmec** **Giorgos Felekis** **Giulio Filippi**

{l.minto.16, jakub.kmec.19, georgios.felekis.19, giulio.filippi.19}@ucl.ac.uk

Abstract

We propose a summarization pipeline to produce abstractive summaries of news articles. We first perform an extractive step at sentence level, in order to *filter* the most relevant sentences in the article. We use this extractive summary as conditioning to fine-tune a GPT-2 model to perform a further abstractive step. Furthermore, we investigate on the shortcomings of the ROUGE metric and propose an alternative summarization evaluation metric and extractor model relying on sentence embeddings. We show that using sentence embeddings similarity measure in the extraction step captures richer latent content and can lead to improved ROUGE scores. Finally, we also show that our pipeline produces coherent and fluent end-to-end summaries.

1 Introduction

The two main paradigms for the task of text summarization are *extractive* and *abstractive*.

Extractive summarization consists in identifying and extracting the most relevant passages (sentences, words or even set of words) from a document. It is a relatively safe approach, that limits to *copying* the source in its most salient points, thus decreasing the chances of inaccuracies and preserving grammaticality.

Abstractive summarization, on the other hand, can rephrase concepts taken from the original document, deviating from the words and forms being used. This approach allows paraphrasing and generalization, which are essential characteristics in high-quality summaries. Nevertheless, modern neural abstractive summarization models (Chopra et al., 2016; Tan et al., 2017), since they rely on recurrent neural network and attention, can be very slow to run and they are known to be prone to bad content selection, leading to factual inaccuracies and redundancies in the generations. This is why there has been a recent shift to using feedforward architectures such as convolutional models and transformers (Vaswani et al., 2017), which address some of the issues. Unlike recurrent architectures, transformers have a logarithmic or constant path length between the output and inputs which makes the gradient flow easier and thus makes learning long term (hundreds to thousands tokens) dependencies easier. They have also been shown to be very good at generating coherent, syntactically and grammatically correct text conditioned on arbitrary input. These are desirable properties which the extractive paradigm lacks and it is thus attractive to try to combine these two approaches.

Finally, there is a hybrid extractive-abstractive architecture (See et al., 2017; Chen and Bansal, 2018; Gehrmann et al., 2018), that aims at combining the advantages of both paradigms, especially better targeted content selection and higher fluency.

We take inspiration from (Subramanian et al., 2019) and propose an extractive-abstractive pipeline, based on the extractive module presented in Chen and Bansal (2018) and the GTP-2 language model (Radford et al., 2019) as the abstractive module, for multi-sentence summarization of news article taken from the CNN/DailyMail dataset (Nallapati et al., 2016b). We select salient content in the articles in the extractive step and feed these extractions to a GTP-like transformer model, which autoregressively generates the final summary conditioned on the extracts. We also propose a variation of the Chen and Bansal (2018) extractive module that makes use of a sentence-embedding-based metric, instead of ROUGE, to create the extraction target labels, especially in the cases where golden summaries are very abstractive in nature.

Finally, we define **pFUSE** and **sFUSE**, two sentence-embedding-based metrics prototypes for evaluating summarization that make use of Universal Sentence Embeddings (Cer et al., 2018). We use these latter metrics to evaluate our extractor variation and compare it with the ROUGE-based one.

We show that using sentence embeddings similarity measure to generate target extraction labels and training on them helps the model capture richer latent content and leads to improved ROUGE scores. Also, as expected, we achieve an increase in FUSE scores compared to the vanilla extractor (Chen and Bansal, 2018). Finally, we also show that our pipeline successfully produces coherent and fluent end-to-end summaries that surpass by 6.24 points the GPT baseline ROUGE-1 scores reported in (Radford et al., 2019).

2 Related work

Extractive Summarization. Recent work on neural extractive summarization has focused on sentence-level extraction (Nallapati et al., 2017, 2016a; Cheng and Lapata, 2016). Nallapati et al. (2016a) propose two RNN based extraction models: a sequential classifier and an out-of-order selector. A greedy selection approach, based off ROUGE scores, is used to select the target labels for both models. Unfortunately, this paper provides extractive baselines only for the DailyMail dataset.

In the same year, Cheng and Lapata (2016) propose an extractive model where attention is used *directly* to select words and sentences from the document representation, an approach parallel to the geometrical reasoning of *Pointer Networks* (Vinyals et al., 2015). A hierarchical document encoder creates the representation both at sentence and document level.

Nallapati et al. (2017) present SummaRuNNer, a GRU based, 2 layers, bidirectional RNN, with a logistic layer that makes binary decisions on sentences based on quantified factors such as information content, salience and novelty. Once again, ground-truths are selected greedily, using ROUGE as scores. This paper also provides extractive baselines for the CNN/DailyMail corpus.

Hybrid Architectures Summarization. Hybrid architectures that mix both extractive and abstractive summarization have proven successful in addressing the poor content selection of abstractive-only models and the limitation of extractive-only models to only reusing words or sentences present in the original document. An hybrid architecture make all the more sense when longer documents are the subject of the summarization task, as is the case for the CNN/DailyMail corpus. Recent work on hybrid architectures (See et al., 2017; Gehrmann et al., 2018; Chen and Bansal, 2018) has proved this approach to be quite successful.

See et al. (2017) propose an hybrid pointer-generator architecture that consist of a *pointer* network that copies words from the original document, in so doing maintaining factual accuracy, and a *generator* network that allows the model to generate new words. A coverage mechanism is put in place at attention-level to avoid repetition in the output.

Gehrmann et al. (2018) address the content selection shortcoming of abstractive-only models by performing an extraction step that selects the most likely phrases to be part of the final summary. These are then fed to an abstractive network that relies on bottom-up attention.

Similarly, Chen and Bansal (2018), from which our baseline extractor model is borrowed, propose a 2-step framework that first selects salient sentences and then compresses and paraphrases them, bridged by an RL system that learns a policy for the number of sentences to be extracted and fed to the abstractive network.

Transformers, GPT and Summarization. Using a transformer language model (GPT-2) for the task of summarization was first proposed in (Radford et al., 2019). They have shown that unconditional language models like GPT are able to implicitly learn to perform tasks such as translation and summarization. By exploiting sequential nature of the training data, the model is able to learn specific tasks with no supervision. For example, the "tl;dr" token can be used to hint the model into generating summaries, which improves performance by 6.4 points. We build on these findings to further fine tune the model for a specific task of summarization.

Sentence Embeddings and Summarization. Traditional research on automatic text summarization is based on metrics as ROUGE, BLEU, METEOR, and others that mainly taking into consideration lexical overlaps. However, in recent years, a new view of content that is based on word or sentence embeddings has motivated numerous studies on summarization evaluation as they are able to capture a higher latent content similarity. Precisely, one of the first and most interesting research findings in this direction is in (Ng and Abrecht, 2015) where they made use of cosine similarity between summary and reference embedding to capture semantic similarities. They show that their method performs better than ROUGE on the baseline datasets as they got better correlations with human judgments. Also, another good example can be found in (Sun and Nenkova, 2019) where they used embedding cosine similarity as a measure of the quality of summarizers on three data sets and make a comparative analysis. Specifically, unlike the previous ones they completely abandon ROUGE and n-gram co-occurrences in the evaluation of semantic similarity and they show that "the maximum value over each dimension of the summary ELMo word embeddings is a good representation that results in high correlation with human ratings".

3 Framework

3.1 Datasets

The main dataset we are going to base our analysis upon is the CNN/DailyMail dataset. The dataset is composed in total of 331,672 article-summary pairs. We use the standard split proposed in (Nallapati et al., 2016b), to measure our model on the same version of the data. The standard split has 286,817 training pairs, 13,368 validation pairs and 11,487 testing pairs. We use the non-anonymized version of this dataset. We chose to work on this dataset because it poses the challenge of having long references (average of 781 tokens) from which to summarize, and also because quite a lot of work has been done on it in the past years and a plethora of baselines are available for us to compare our method with.

3.2 Extractor

The extractor model takes in as input a document and recurrently outputs a hard-list of salient sentences from the doc. We use the same vanilla extractor architecture as used in Chen and Bansal (2018). This architecture comprises of two stages: first, a hierarchical neural model to learn the sentence-level representations, and second, a selection network that selects the most salient sentences given the representations computed by the previous layer. In turn, the first layer includes a temporal convolutional model that computes the sentence representation r_j, and a bidirectional LSTM-RNN to incorporate context and dependencies from both directions, so that a sentence is not considered in a vacuum but in the larger context of the doc. The selection network used in Chen and Bansal (2018) is an LSTM-RNN *Pointer Network* (Vinyals et al., 2015).

Target labels for the extractor are created via ROUGE-L scores. For each sentence in the abstract a_i, the sentence in the document d_j with the highest ROUGE-L score ROUGE-L(a_i, d_j) is selected as a target. One sentence in the document is selected for each sentence in the abstract. The probability of extracting already extracted sentences is forced to zero to prevent redundancy.

We present now a variation of this model that bases target label creation off a different type of metric.

3.3 Sentence-Embedding-based Extractor

The extractor model presented in Chen and Bansal (2018) is trained on target labels computing the ROUGE-L between article and summary sentences. Similarly to Nallapati et al. (2017), for each sentence s_j in the reference summary a sentence d_j is selected in the article such that:

$$d_j = argmax_i(ROUGE_{L,recall}(d_i, s_j))$$

We observe that CNN/DailyMail dataset golden summaries are quite abstractive in nature, that is, they reformulate concepts from the original articles quite concisely by making extensive use of paraphrasing and unstated context. For this reason, we wish to use a target label selection metric that does not rely as much on the lexical overlapping, as these alone couldn't very well capture paraphrasing and generalizations, but rather on the semantic similarity between two sentences.

In a different scenario, where ground truth summaries were not as abstractive, but merely short repetitions of the words used in the original document, it wouldn't have made as much sense to use a semantic similarity metric, as the lexical overlapping would have been more than enough to create good labels.

We compute sentence-level semantic similarity as the inner product of the embeddings of the sentences in the article and the abstract. To compute the embeddings we rely on the Universal Sentence Embeddings model (Cer et al., 2018). More specifically, for each sentence s_j in the reference summary, we extract the sentence d_j

$$d_j = argmax_i(\mathbf{e}(s_j) \cdot \mathbf{e}(d_i))$$

where $\mathbf{e}(\cdot)$ is the embedding function of a sentence.
We then train the extractor model in the same way we trained it with the ROUGE-based labels.

3.3.1 Hypothesis Testing

We wish to test the hypothesis that sentence embedding similarity yields *better* extractions than n-grams overlapping (ROUGE). We compute the extraction ground truths first using ROUGE-L, as in (Subramanian et al., 2019; Chen and Bansal, 2018), and then using our proposed sentence embedding based metric. In both cases, for each sentence in the golden summary we find the sentence in the article with, respectively, the highest ROUGE-L and SES score and we store their indexes. For each sentence that has different extraction candidates under ROUGE-L and SES we ask an human evaluator to decide which candidate article sentence best represents the reference sentence in the golden summary. Options are presented in a random order to prevent any bias and a third option is available in case the two candidates are both right or both wrong. The evaluator is also asked to be conservative in making his decisions, so that only if an extraction candidate is *clearly* better than the other it gets selected.

3.3.2 Examples

In Table 1 we show an example of the tests the human evaluator was prompted with during the metric validation experiment. As said before, options are displayed in a random order, to avoid bias in the selection. An analysis on the extraction labels produced by ROUGE and SES together with some significant examples are reported later on in the paper, under the Analysis section.

3.3.3 Statistics

Out of 120 examples compared by an human evaluator, for 69 of them SES was judged as better than ROUGE. The question is: is this a statistically significant test or is it just a statistical fluke?

Reference #1
scientists : an epic migration by jupiter led to destruction of “ super-earths ”.
Option 1
two scientists are suggesting that the inner solar system once played host to a bunch of “ super-earths ” – planets that were larger than our own but smaller than neptune .
Option 2
he and his co-author – gregory laughlin of university of california , santa cruz – are building on a scenario of jupiter ’s migration that was previously put forward by other scientists .
Option 3
Can’t tell, both right or both wrong.

Table 1: A sample of test prompted to the human evaluator.

We will model our experiment as a sequence of independent binary random variables $X_1, , X_{120}$ where for each i, $X_i = 1$ if ROUGE is better than SES on trial i and $X_i = 0$ otherwise. We denote by p the probability under a random trial that ROUGE is better than SES, in equations

$$P(X_i = 1) = p$$

The null hypothesis and alternative hypothesis denoted respectively H_0 and H_1 are

$$H_0 : p \geq \frac{1}{2}$$

$$H_1 : p < \frac{1}{2}$$

Since the variance of our data is unknown, we must perform a one sided t-test. To do so we use the statistic

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

where $\bar{X}$ is the sample average

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{51}{120}$$

s is the sample standard deviation

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{1}{\sqrt{n}}\sqrt{69\bar{X}^2 + 51(1-\bar{X})^2}$$

And μ is the hypothesized mean, which we take to be 0.5 ie the left-most value of our null Hypothesized set. With these values we compute the statistic to be

$$T = \frac{120(\bar{X}-\mu)}{\sqrt{69\bar{X}^2 + 51(1-\bar{X})^2}} \approx -1.763$$

We know that T is distributed as a t-distribution with $n - 1 = 119$ degrees of freedom. So we can calculate the probability of getting a T-value as extreme as -1.763 under the null hypothesis by using the cdf of a t_{119} distribution.

$$P_{\mu=0.5}(T \leq -1.763) = 0.0403$$

So with probability greater than 95%, the null hypothesis is False.

This concludes the proof of statistical significance of our test. There are of course many problems with this analysis involving : judgement of tester, potential non-independence of trials, the set of all texts not being a measurable set, etc. However this is a good preliminary result indicating that there is potential in SES as a method for comparing sentences for summarization.

3.4 Abstractor

The abstractive step makes use of a decoder only transformer architecture identical to the GPT-2 language model (Radford et al., 2019), which is fine-tuned to our summarization task. This model was trained using a causal language model objective and is powerful in predicting the next token in a sequence. We use this property to autoregressively generate full summaries, conditioned on the input context. The input context, in raw text sentences format, comes from the extractive step and is encoded using a Byte-level Byte-pair tokenizer (Sennrich et al., 2015). The output vocabulary has a size of 50259, which includes two special tokens, <TLDR >and <EOD >. These are meant to signal the beginning and end of the summary respectively.

As a starting point we used the OpenAI GPT-2 English model which has 12-layers, 768-dimensional embeddings, 12-heads and a total of 117M parameters. We build on the HuggingFace transformers (Wolf et al., 2019) PyTorch implementation of GPT-2. The only change is in the way we calculate loss during training. We assume that the model has already learned how to generate syntactically and grammatically correct text and we thus only calculate the loss with respect to the summary (all tokens that come after <TLDR >). This should intuitively reduce noise in the feedback the model receives at each update.

3.5 Training

We fit the extractor and abstractor separately. The extractor is trained on the original data from the CNN/DM corpus, while the abstractor is trained on the ground truth extractions paired with their respective summary.

One of the main limitations of our pipeline is that, as is, it cannot be trained end-to-end. We, therefore, have to set fixed the number of sentences to be extracted, which will sometimes result in the extraction of irrelevant sentences and other times in the loss of valuable information. We set the number of extractions to maximize ROUGE scores. Also, since the abstractor is trained on the oracle extractions, it might struggle to handle input noise well at prediction time.

Our computational resources were very limited, which is why we trained the transformer model on a random sample of 200k examples for 3 epochs, which took 10 hours on a Tesla p100 GPU. Based on Figure 1, the evaluation loss has still not fully converged so there is potential to imporve the results further by investing more computational resources.

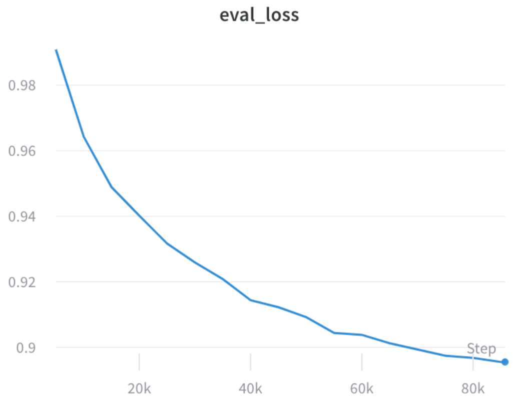


Figure 1: Evaluation loss after 85000 steps (batch=7)

4 Evaluation methods and metrics

The most commonly used way to evaluate the validity of automatically produced summaries is to compare them with human referenced model summaries. Evaluation techniques can be divided into intrinsic and extrinsic. An intrinsic evaluation tests the summaries in and of themselves on how informative and coherent they are. On the other hand an extrinsic evaluation tests the impact of the summarization on the completion of a bigger task. In our case we are following an intrinsic based on content evaluation procedure. The main drawback of the evaluation systems in text summarization is that most of the time a reference summary is needed, and for some methods more than one, to be able to compare automatic summaries with models. Hence, text summarization evaluation involves human judgments of different quality metrics, for example, coherence, conciseness, grammatically, readability, and content (Mani, 2001). In any case, evaluation methods need a set of summaries as an input to set the standards for two subgroups based on evaluation metrics: text quality and content evaluation and a set of automatic summaries.

4.1 ROUGE score metric

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation and is a package for automatic evaluation of summaries (Lin, 2004). ROUGE is based on n-grams to measure the similarity between a summary generated by a model and a reference summary of human and is recall-oriented e.g. how much the n-grams in the reference summary appeared in the model generated summary. Here we briefly present the two basic ROUGE metrics : ROUGE-N and ROUGE-L. In our experiments we used ROUGE-L and ROUGE-N (ROUGE-1, ROUGE-2).

Following the notation of (Lin, 2004) we have:

4.1.1 ROUGE-N

ROUGE-N

$$= \frac{\sum\limits_{S \in RS} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in RS} \sum\limits_{gram_n \in S} Count(gram_n)}$$

where RS are the reference summaries, $gram_n$, and $Count_{match}(gram_n)$ are the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. Hence, ROUGE-N calculates the overlap of n-grams between the model and reference summaries. Specifically, ROUGE-1 calculates the overlap of unigrams (words) where ROUGE-2 calculates the overlap of bigrams between the two summaries.

4.1.2 ROUGE-L

Suppose we have a sequence $X = [x_1, x_2, ..., x_m]$. A subsequence of X is a sequence $Y = [y_1, y_2, ..., y_n]$ if there exists a strict increasing sequence $I = [i_1, i_2, ..., i_k]$ of indices of X such that for all $j = 1, 2, ..., k$, we have $x_{i_j} = y_j$ (Cormen et al., 2001). The longest common subsequence (LCS) of two sequences X and Y is a common subsequence with maximum length. ROUGE-L measures the longest matching sequence of words using LCS and it does not require consecutive matches but in-sequence matches that reflect sentence level word order.

4.2 Sentence Embeddings Based Metric

4.2.1 Motivation

In the recent years, and because there has been no profound improvement in performance on the main datasets in the field of text summarization a lot of concern has arised on ROUGE and how naturally this describes a human's idea of an optimal summary (Schluter, 2017). Specifically, because as mentioned above, ROUGE takes into account overlaps of n-grams, many times is unable to capture a latent semantic similarity between summaries, especially in the most abstractive ones.

In this report we propose a new metric for automatic evaluation of summaries based on similarity scores using embeddings from the universal sentence encoder. We will then use this new metric to measure the performance of the ROUGE and SES based extractors.

4.2.2 Universal Sentence Encoder

The Universal Sentence Encoder (USE) by Google (Cer et al., 2018) is a publicly available model for encoding sentences into embedding vectors and provide strong transfer performance on a number of NLP tasks. The original paper shows that these sentence embeddings capture a richer semantic content as opposed to combining individual word embeddings.

4.2.3 FUSE metric

We introduce a new sentence embedding based evaluation metric for text summarisation called **FUSE**. The main idea behind FUSE is that based on the sentence embeddings of the Universal

Sentence Encoder it calculates the similarity of two summaries by calculating the inner product of their embedding vectors. Specifically, we present two version of FUSE:

I. A **paragraph-based** one (p-FUSE) where we calculate the embedding $e(ref)$ of the reference summary and the embedding $e(gen)$ of the generated one and calculates the inner product of them in order to produce a final score. There is no hard limit on how long the paragraph is but roughly, "the longer, the more 'diluted' the embedding will be."

II. A **sentence-based** one (s-FUSE) follows: Suppose we have a reference summary with K sentences $R_1, R_2, ..., R_K$ and a model-generated summary with L sentences $G_1, G_2, ..., G_L$. Firstly, we define a score for each sentence in the reference summary as follows:

$$score(R_k) = max\{e(R_k) \cdot e(G_l) | \forall l \in 1, ..., L\}$$

where $e(\cdot)$ is the embedding function of a sentence given by the Universal Sentence Encoder. Similarly, we can define a score for each sentence in the reference summary:

$$score(G_l) = max\{e(G_l) \cdot e(R_k) | \forall k \in 1, ..., K\}$$

Both on $score(R_k)$ and $score(G_l)$ we demand an injective property, meaning that there should be a 1-1 assignments between the two sentences sets (i.e .a sentence of the generated summary can only be selected once as the maximum for a sentence of the reference one and vice versa). However if the size of the summaries is different we keep the sentences with the best scores and assign score equal to zero to the others.

Now we can define two new similarity measures(sim_R, sim_G) between two summaries R and G as below:

$$sim_R(R, G) = \frac{1}{K} \sum_{k=1}^{K} score(R_k)$$

and

$$sim_G(R, G) = \frac{1}{L} \sum_{l=1}^{L} score(G_l)$$

We can easily see that is reasonable to think R-similarity as a recall measure (R) and G-similarity as a precision (P) one. Hence we can define a F_1-score:

$$F_1 = 2 * \frac{P}{P + R}$$

which is our final FUSE score. So, overall for two summaries R,G we have:

$$sFUSE(R, G) =$$

$$\frac{\frac{1}{KL} \sum_{k=1}^{K} score(R_k) \sum_{l=1}^{L} score(G_l)}{\frac{1}{K} \sum_{k=1}^{K} score(R_k) + \frac{1}{L} \sum_{l=1}^{L} score(G_l)}$$

Finding the optimal match for each sentence in the document and, viceversa, for each sentence in the abstract would be computationally expensive. That's why we have opted for a greedy approach when computing sFUSE.

5 Results

We are going to report our results in two separate tables: one focusing on extractive results, another focusing on abstractive results. In all three tables, we are going to report scores for two separate pipelines: one based on the standard ROUGE extractions, and another relying on the SES extractions.

5.1 Extractive Summarization

In Table 2 we report both the ROUGE and FUSE scores achieved by our extraction models with number of extractions fixed to 3 and 4. We also include in the table scores for our baseline model (Chen and Bansal, 2018), with ROUGE based extractions and with number of extractions also fixed to 3 and 4, the lead-3 baseline from (See et al., 2017) with relative computed pFUSE and sFUSE scores, the extractive model scores reported in (Nallapati et al., 2017), and, finally, to put all the scores in perspective, the extractive SOTA scores from (Liu and Lapata, 2019).

As can be seen in Table 1, at prediction time, our extractor model, compared to (Chen and Bansal, 2018) extractor, not only gets better FUSE scores, as expected, but also gets higher ROUGE-1, ROUGE-2 and ROUGE-L scores. We explain this collateral effect partially as result of the correlation between words overlapping and semantic similarity, but also, potentially, as result of easier disentanglement of the sentence representations under these different labels. We also manage to surpass, even if ever-so-slightly, the lead-3 baseline as computed in (See et al., 2017). Nevertheless, our model significantly underperforms the state of the art.

5.2 Abstractive Summarization

Table 3 shows the results achieved by different abstractive approaches as well as lede-3 and random-3 taken from Radford et al. (2019). GPT-2 (zero) is a zero shot language generation with the article used as context. GPT-2 TL;DR is also zero shot but with the additional token used as a hint. Our apporach, which consists of an extractive step and fine tuned GPT-2 model outpreforms all of these baselines. Lede-3 is still significantly better in all metrics.

MODEL	ROUGE		
	1	2	L
Ext+GPT-2 oracle	47.93	25.25	38.05
Lead-3	40.34	17.70	36.57
Ext+GPT-2 tuned	**35.58**	**14.69**	**30.68**
Seq2seq + Attn	31.33	11.81	28.83
GPT-2 TL;DR	29.34	8.27	26.58
Random-3	28.78	8.63	25.52
GPT-2 zero	21.58	4.03	19.47

Table 3: Abstractive results.

6 Analysis

6.1 Abstractivness

Despite not performing as well as pure extraction in the ROUGE metrics, the abstractive approach produces much more coherent summaries, which often do a better job at capturing the semantics of the article. This can be seen in Example 2 in Table 4. Since one of the main advantages of abstractive approaches is their ability to rephrase the input context, it is expected the ROUGE L metrics (which capture word overlap) will not be as high. Example 4 nicely shows the model's ability to rephrase input text into shorter, yet more coherent sentences. It breaks the first long sentence down into two simple ones, which are clearer and easier to digest.

The model also learns to do effective sentence- and word-level extraction and eliminates unnecessary words while preserving fluency and information content as shown in Example 1. The generated output includes all important facts (nationality, airport, dates...), but reduces the number of words in each sentence.

The biggest disadvantage of this approach is that factual information is not guaranteed to be preserved and it is possible the model will suffer from bias in the data as shown in Example 3.

We also noticed that the model is prone to generate new, made-up content which isn't part of the initial article. Our hypothesis is that this is because we use a pre-trained GPT-2 model that was trained on external data that contains affine subjects and can therefore create a bias in generation.

6.2 SES vs ROUGE: Round 2

In Table 5 we show a partial and a full sample extraction from SES and ROUGE. We notice that one of SES shortcomings is matching sentences that are very entity heavy, that is include names of people, places or companies. We think this is because such entities have little to no contribution to the sentence embeddings and cannot be therefore captured in the similarity. An example of this can be found in A.1 abstract #1 extractions.

On the other hand, partial abstract #2 is an interesting example of the level of abstractivness SES is able to detect. Rouge is lured to its extraction by the "*1,500 jews in esfahan*" overlap, but it ends up selecting a sentence that has nothing to do with reference's true meaning. SES, instead, captures the indirect meaning of the quote and successfully finds the semantic overlap with the reference sentence.

In Appendix A.1 we present a heat map visualization of the scores produced by ROUGE and SES at target selection time. Specifically, we can observe how SES offers a bigger uncertainty about which is the best sentence to extract as it captures a higher latent semantic correspondence of the sentences. On the other hand, ROUGE seems to be quite confident about the choice of the sentence as it only takes into consideration overlaps of n-grams, even though many times it makes a worse choice as we saw previously.

[0] All ROUGE scores are computed through the official ROUGE 155 package.

Models	ROUGE-1	ROUGE-2	ROUGE-L	pFUSE	sFUSE
Chen and Bansal (2018) ext (3)	39.47	17.89	35.83	69.05	42.89
Chen and Bansal (2018) ext (4)	36.91	17.01	33.94	69.68	47.31
SE-ext (3)	**40.36**	**18.07**	**36.74**	69.31	43.68
SE-ext (4)	38.27	17.20	35.22	**69.71**	**47.78**
lead-3 (See et al., 2017)	40.34	17.70	36.57	68.50	42.26
SummaRuNNer (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-
(Liu and Lapata, 2019) SOTA	43.85	20.34	39.90	-	-

Table 2: ROUGE and FUSE scores for extractive summarization on the CNN/Dailymail dataset. Models with * sign were tested on the anonymized version of the dataset and so they are not entirely comparable. We also include the state of the art BERT based summarization model scores as a reference.

7 Conclusion and Future Work

In this paper we have proposed a summarization pipeline to produce abstractive summaries of news articles end-to-end. We have shown that our model can summarize long texts fluently and coherently. Nevertheless, our model does not match the state of the art in ROUGE-1,2,L scores. We also argue in favour of a new comparison metric based on Universal Sentence Embeddings (USE) to replace ROUGE scores, showing that it is a promising alternative capturing richer and more abstract similarities between summaries.

The most straightforward way to improve the current performance of our model would be by training on the full CNN/DM corpus and for more epochs, while so far, because of time and computational constraints, we have only trained on 200k samples for 3 epochs. Using a larger GPT-2 model, with a wider input window could also potentially lead to major improvements.

To address extraction size a bridging reinforcement learning system could be used to learn a policy for the optimal number of sentences to be extracted. Perhaps, even taking into consideration the abstractor input window size.

Finally, we proved ROUGE to be not as effective as SES at capturing semantic similarity between sentence. In turn, we notice SES struggles with entity heavy sentences, that is sentences with names of people, place or companies. A combination of ROUGE and SES could be used to build a more complete metric and address both the aforesaid shortcomings.

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *CoRR*, abs/1805.11080.

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.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2001. *Introduction to Algorithms*, 2nd edition. The MIT Press.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. *CoRR*, abs/1808.10792.

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.

Inderjeet Mani. 2001. *Automatic Summarization*, volume 3.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. *CoRR*, abs/1611.04230.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016a. Classify or select: Neural architectures for extractive document summarization. *CoRR*, abs/1611.04244.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher Pal. 2019. On extractive and abstractive neural document summarization with transformer language models.

Simeng Sun and Ani Nenkova. 2019. The feasibility of embedding based automatic evaluation for single document summarization. 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 1216–1221, Hong Kong, China. Association for Computational Linguistics.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks.

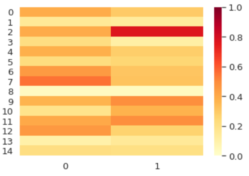
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

A Appendices

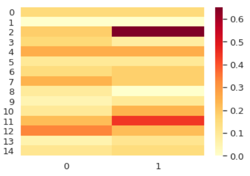
A.1 Extraction Heatmaps

The following heatmaps produced from the next example of the CNN/DailyMail dataset. The numbers represent the order of the sentences on the article and the summary.

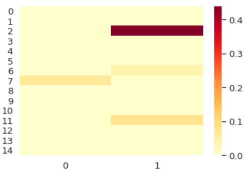
Article:



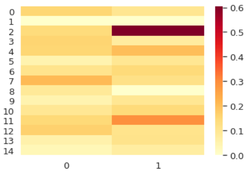
(a) FUSE



(b) ROUGE-1



(c) ROUGE-2



(d) ROUGE-L

(0)"-lrb- cnn -rrb- the only thing crazier than a guy in snowbound massachusetts boxing up the powdery white stuff and offering it for sale online ?"(1)"people are actually buying it ."(2)"for $ 89 , self-styled entrepreneur kyle waring will ship you 6 pounds of boston-area snow in an insulated styrofoam box – enough for 10 to 15 snowballs , he says ."(3) "but not if you live in new england or surrounding states .... (6) "with more than 45 total inches , boston has set a record this winter for the snowiest month in its history . most residents see the huge piles of snow choking their yards and sidewalks as a nuisance , but waring saw an opportunity ."(7) "according to boston.com , it all started a few weeks ago , when waring and his wife were shoveling deep snow from their yard in manchester-by-the-sea , a coastal suburb north of boston . he joked about shipping the stuff to friends and family in warmer states , and an idea was born ."(8) "his business slogan : " our nightmare is your dream ! """.....(12) "many of his customers appear to be companies in warm-weather states who are buying the snow as a gag , he said ."(13) "whether waring can sustain his gimmicky venture into the spring remains to be seen . but he has no shortage of product ."(14) "" at this rate , it 's going to be july until the snow melts , " he told boston.com . " but i 've thought about taking this idea and running with it for other seasonal items . maybe i 'll ship some fall foliage ."

Summary:

(0)"a man in suburban boston is selling snow online to customers in warmer states ."(1)"for $ 89 , he will ship 6 pounds of snow in an insulated styrofoam box ."

	Extract	Generated summary	Reference summary
1	”yahya rashid , a uk national from northwest london , was detained at luton airport on tuesday after he arrived on a flight from istanbul , police said .’s been charged with engaging in conduct in preparation of acts of terrorism , and with engaging in conduct with the intention of assisting others to commit acts of terrorism . both charges relate to the period between november 1 and march 31 . rashid is due to appear in westminster magistrates ’ court on wednesday , police said .”	” uk national from london was detained at luton airport on tuesday. he is charged with engaging in conduct in preparation of acts of terrorism. charges relate to period between november 1 and march 31. ”	”london ’s metropolitan police say the man was arrested at luton airport after landing on a flight from istanbul . he ’s been charged with terror offenses allegedly committed since the start of november .”
2	”-lrb- cnn -rrb- since iran ’s islamic revolution in 1979 , women have been barred from attending most sports events involving men .but the situation appears set to improve in the coming months after a top iranian sports official said that the ban will be lifted for some events .a plan to allow “ women and families ” to enter sports stadiums will come into effect in the next year , deputy sports minister abdolhamid ahmadi said saturday , according to state-run media .”	” iran’s islamic revolution in 1979 barred women from attending most sports events involving men. a top iranian sports official said the ban will be lifted for some events. ”	”iranian sports official : the ban will be lifted for some events in the coming year .but he says “ families are not interested in attending ” some sports matches .”
3	”here ’s a journey from the vatican to the holy land , around the moon and the earth ’s tilting axis , to germany and the united states to try to explain the complex holiday called easter . and you ’ll learn to how to color easter eggs with kool-aid .jesus rode on the back of a donkey into jerusalem on palm sunday where crowds celebrated him as the messiah and laid down palm branches in his path . but the crowd and the romans turned on him in the course of the week , according to the bible , leading to his crucifixion and resurrection .at the vatican , holy week began with palm sunday last week and culminated in easter sunday mass with multiple celebrations in between to mark the final week of christ ’s mortal life .”	” jesus rode on the back of a donkey into jerusalem on palm sunday. crowds celebrated him as the messiah and laid down palm branches in his path. but the crowd and the romans turned on him in the course of the week. ”	”easter is a key event in the christian faith , but where did the easter bunny come from ? why is the date different every year , and what does it have to do with the moon ?”
4	”hong kong -lrb- cnn -rrb- six people were hurt after an explosion at a controversial chemical plant in china ’s southeastern fujian province sparked a huge fire , provincial authorities told state media . the plant , located in zhangzhou city , produces paraxylene -lrb- px -rrb- , a reportedly carcinogenic chemical used in the production of polyester films and fabrics .the blast occurred at an oil storage facility monday night after an oil leak , though local media has not reported any toxic chemical spill .”	” six people hurt after explosion at chemical plant in fujian province. explosion sparked huge fire, provincial authorities say. blast occurred at an oil storage facility monday night. ”	”a blast rocks a chemical plant in china ’s southeastern fujian province for the second time in two years .six were injured after the explosion and are being hospitalized .the explosion was triggered by an oil leak , though local media has not reported any toxic chemical spills .”

Table 4: Examples of abstracted summaries.

Abstract #1
(1) jordan spieth holds lead in 2015 masters. **(2)** strong starts from mcilroy and woods. **(3)** both fall away as 21 year old spieth takes control.
ROUGE extraction
(1) saturday at the masters, like any pga tournament, has been dubbed ’moving day’ . **(2)** rory mcilroy went out in 32 and briefly raised the crowd’s hopes that he had a sniff of completing an improbable grand slam on sunday night. **(3)** spieth’s 15 birdies are just 10 away from phil mickelson’s masters mark set in 2001. he could also break tiger woods 270 set in 1997.
SES extraction
(1) cnn saturday at the masters, like any pga tournament, has been dubbed ’moving day’ . **(2)** players rose and players fell away on moving day at the 2015 masters. **(3)** when reminded of some of the great augusta comebacks, including nick faldo’s 11 shot swing in 1996, tiger woods still believes anything is possible.

Partial #2
1,500 jews call esfahan home despite tensions between iran and israel.
ROUGE extraction
there are about 1,500 jews in esfahan these days. the community ’s leaders conduct religious studies for the younger members of the congregation.
SES extraction
”israel and iran are countries,”he said.”and we consider ourselves iranian jews,not israeli jews.so the hostilities between israel and iran do not affect us.”

Table 5: Examples of ROUGE and SES based extractions. The second example is a partial extraction