Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1796–1802
Marseille, 11–16 May 2020

Evaluation Metrics for Headline Generation Using Deep Pre-Trained Embeddings

Abdul Moeed†, Yang An†, Gerhard Hagerer†, Georg Groh
Research Group Social Computing, Department of Informatics
Technical University of Munich, Germany
{abd.moeed, yang.an, gerhard.hagerer}@tum.de, grohg@in.tum.de
† These are equally first authors and appear in random order.

Abstract

With the explosive growth in textual data, it is becoming increasingly important to summarize text automatically. Recently, generative language models have shown promise in abstractive text summarization tasks. Since these models rephrase text and thus use similar but different words as found in the summarized text, existing metrics such as ROUGE that use n-gram overlap may not be optimal. Therefore we evaluate two embedding-based evaluation metrics that are applicable to abstractive summarization: Fréchet embedding distance, which has been introduced recently, and angular embedding similarity, which is our proposed metric. To demonstrate the utility of both metrics, we analyze the headline generation capacity of two state-of-the-art language models: GPT-2 and ULMFiT. In particular AES shows close relation with human judgments in our experiments and has overall better correlations with them compared to ROUGE. To provide reproducibility, the source code plus human assessments of our experiments is available on GitHub[1].

Keywords: Evaluation Methodologies, Language Modelling, Natural Language Generation, Summarization, Textual Entailment and Paraphrasing, Statistical and Machine Learning Methods

1. Introduction

The recent development of generative language models (LMs) is leading to new capabilities regarding the quality of text generation (Radford et al., 2019). This also holds true for tasks such as abstractive summarization, which is related to the language model generating summaries de nouveau and paraphrasing the text in its own words (Moratanch and Chitrakala, 2016). This is of high importance considering the large and always increasing amount of available texts and their relevance for humans.

An advantage of abstractive summarization is its superior readability (Hsu et al., 2018) compared to extractive summarization where keywords from the text are extracted and rearranged (Lin and Hovy, 2003). This benefit can be used for generating realistic headlines (Takase et al., 2016; See et al., 2017). However, it remains a challenge to find a faithful evaluation metric. ROUGE (Lin, 2004) - a standard performance metric for extractive summarization - is not always ideal for abstractive summarization, since readability is not taken into account (Paulus et al., 2017). Instead, it only accounts for n-gram overlap which is a problem for use cases when summaries rephrase the respective content using different but similar words.

To address this problem, pre-trained semantic similarity embeddings such as InferSent (Conneau et al., 2017) have been used successfully to evaluate the quality of GAN-based text generation (Semeniuta et al., 2018). Therefore, the concept of the Fréchet distance (Heusel et al., 2017), which is a well-known procedure for computer vision, is successfully applied for text generation as well. Due to the novelty of the approach, it appears unclear *how this method relates to human judgment on the task of abstractive summarization*. Further, it stays unclear how the concept works with more recent pre-trained embeddings than InferSent, and based on which language models these research questions could be solved.

Since most recent pre-trained embedding models are trained on sentences, we perform headline generation as an instance for abstractive summarization in order to evaluate the general feasibility of the approach. More specifically, we generate headlines for user product reviews and news stories using OpenAI's GPT-2 (Radford et al., 2019) after comparing its performance to fastai's ULMFiT (Howard and Ruder, 2018). For a comprehensive analysis, GPT-2 is trained on four datasets: a sub-dataset of the Amazon Product Dataset (He and McAuley, 2016; McAuley et al., 2015), CNN/Daily Mail (Hermann et al., 2015), Newsroom (Grusky et al., 2018) and Gigaword (Napoles et al., 2012) [2].

In order to generate headlines, we fine-tune and condition the language models. Based on the Universal Sentence Encoder (USE) (Cer et al., 2018) we derive Fréchet distances and depict their relation with human judgments. Additionally, we show another measurement based on angular similarity (Cer et al., 2018) with similar properties as Fréchet distance for the evaluation of generated headlines.

2. Methodology

2.1. Language Models

To generate summaries, we use two autoregressive language models: GPT-2 and ULMFiT. BERT (Devlin et al., 2018) is another powerful language model, and has been previously modified and used for extractive summarization (Liu, 2019). However, owing to its bi-directional nature, BERT expectedly performs poorly with masked input for text generation, and is thus not further considered. Released in February 2019, OpenAI's GPT-2 achieved state-of-the-

[1] https://github.com/Abdul-Moeed/headline-gen-metrics

[2] Gigaword corpus is taken from `https://github.com/harvardnlp/sent-summary`

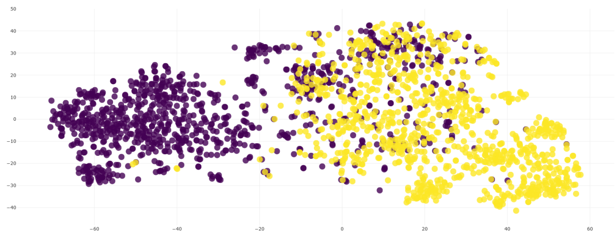

Figure 1: Dimension reduction of embeddings for Musical instruments (purple) and patio, lawn and garden (yellow) using PCA and t-SNE.

art performance on a variety of tasks in the zero-shot setting. Furthermore, it is trained in an unsupervised regime, with no domain-specific knowledge. The model is trained on OpenAI's custom "WebText" dataset. For our task, we fine-tune the smallest model available, with about 117 million parameters. Universal Language Model Fine-tuning for Text Classification (ULMFiT) was introduced by fastai in 2018, and is still used for many NLP tasks, including text generation. It uses cyclical learning rates (Smith, 2015) to converge faster compared to other models.

2.2. Automatic Evaluation Metrics

2.2.1. ROUGE

The current standard for evaluating summaries is ROUGE. Most authors report their ROUGE-1, ROUGE-2 and ROUGE-L scores as the only automatic evaluation score besides human evaluation (Liu, 2019; Nallapati et al., 2016; Nallapati et al., 2017; Paulus et al., 2017). Because they only compare n-gram overlap, ROUGE scores are agnostic of semantic similarity between reference and hypothesis summaries. This property is magnified in abstractive summarization, as the models try to paraphrase the original content. Thus they might use different wording compared to a reference text whereas the semantic meaning stays identical. The present paper addresses the issue by using embedding-based metrics. BLEU (Papineni et al., 2002) – a standard metric for machine translation – is also sometimes used for text summarization (Graham, 2015) though it is far from being as ubiquitous as ROUGE, and is known to not perform any better (Graham, 2015).

2.2.2. Proposed Embedding-Based Metrics

In this paper we use metrics based on sentence embeddings. The intuitive idea is to embed the generated and reference headline and the corresponding article text, then compare the semantic similarity between each of them.

Embedding Model As features for semantic similarity of sentences, we use the pre-trained embedding module Universal Sentence Encoder (USE) by Google (Cer et al., 2018). The model takes text of variable length as input, i.e. word sequences, sentences or small paragraphs, and encodes it as a 512 dimensional vector which can be used for text classification, semantic analysis, or other natural language processing tasks. For this study, we use the USE version based on the deep averaging network architecture (DAN) (Iyyer et al., 2015), which is available for download[3]. Each word in the sentence is first mapped to a word2vec (Mikolov et al., 2013) embedding before the USE averages them. This vector representation is then pushed through a feed forward neural network which produces a normalized sentence embedding. The DAN was pre-trained by the authors on English data from Wikipedia articles, question answering web pages, web news and discussion forums. The corresponding training tasks included conversation response prediction, quick thought, natural language inference (NLI) and translation ranking. The model is thus pre-trained to resemble semantically meaningful feature vectors which are suitable for a wide range of tasks. Consequentially, we utilize the pre-trained model as it is provided to calculate feature vectors for headlines and respective stories.

Review: My Weber grill cover couldn't make it 2 years in the Chicago winter and summer. [...] It has survived a few normal rain/thunderstorms without blowing off and protected the grill.I'll update my review in the future.[Update April, 2014]The cover survived a harsh cold winter, and more importantly my grill survived.
Reference: A replacement for a weber genesis
Generated 1: My grill survived the Midwest's toughest winter.
Generated 2: Still works today because of superior quality and durability.
Generated 3: Nice cover but nothing survives frost.

Table 1: Example headlines generated by GPT-2 on the 'patio, lawn and garden' subset.

Embeddings from semantically similar content ideally lie close to each other. We validate this by first applying PCA to reduce the dimensionality of the embedding space and then feeding the result to t-SNE to project the vector onto two dimensions for visualization (Figure 1). An interactive version can be found in our repository. Generated headlines from the 'musical instruments' sub-dataset (purple) on the one hand and 'patio, lawn, and garden' sub-dataset (yellow) on the other hand are separated adequately. Digging further, we observe that headlines from the same product lie close together, e.g. headlines of effect pedal reviews lie in the upper left corner. Moreover, non-informative headlines, e.g. "great product", "cheaply made", are centered between both clusters, as they do not contain identifying information about the product being reviewed. We conjecture that omitting the centered headlines during training would produce more reasonable headlines.

Angular Embedding Similarity Cer et al. (2018), the authors of USE, propose angular similarity to compare the semantics of two embeddings as

$$\operatorname{sim}(\mathbf{u}, \mathbf{v}) = 1 - \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|u\| \, \|v\|}\right) \cdot \frac{1}{\pi} \qquad (1)$$

which is a modification of cosine similarity to perform better on small angles.

To our knowledge, USE with angular similarity between generated and reference samples has never been used as an

[3] https://tfhub.dev/google/universal-sentence-encoder/2

Language Model	AES	Human
ULMFiT	0.575	17.7%
GPT-2	0.623	53.3%

Table 2: Average similarity of headlines generated by our fine-tuned language models compared to the corresponding reviews from Amazon. AES is our proposed metric defined in formula 2. Human is the similarity as perceived by two human annotators. The higher the values, the better is the headline generation capability of the language model.

evaluation metric for headlines or other kinds of abstractive summaries before. Therefore we propose the *angular embedding similarity* (AES) as the average of all angular similarities between the USE embeddings of two related text samples. For instance, when comparing all stories $\in$ R and their corresponding headlines $\in$ H of a corpus, the AES between them is defined as

$$\text{AES}_{S,H} = \frac{1}{n} \cdot \sum_{i=1}^{n} \text{sim}(\hat{s}_i, \hat{h}_i), \tag{2}$$

with $\hat{s}_i$ and $\hat{h}_i$ being the USE embedding of s_i and h_i, i.e., the ith story and corresponding headline, and n the total number of stories in that corpus.

As described in the experiments section, we evaluate AES between reference headlines and stories, generated headlines and stories, and generated headlines and reference headlines.

Fréchet Embedding Distance Fréchet distance (Fréchet, 1957) is a measurement to compare two Gaussian distributions

$$\text{FID}(r, g) = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{0.5}\right) \tag{3}$$

where r refers to the reference sample distribution and g to the generated sample distribution, μ and Σ to their corresponding means and covariance matrices.

The Fréchet distance has already been used successfully in computer vision to evaluate generative models (Heusel et al., 2017). Recently, it has been used in natural language generation as an alternative to ROUGE (Semeniuta et al., 2018). There, InferSent is utilized (Conneau et al., 2017) to compare the output of GANs for language generation. d'Autume et al. (2019) later calculated the Fréchet distance on USE embeddings and called it the Fréchet embedding distance (FED). The authors also noticed a drawback of FED: its sensitivity to length. We also confirm this observation experimentally in section 4. In contrast to AES, FED is a distance and lower scores mean higher similarity. Also worth noting is the fact that FED requires multiple samples for computing the distance between distributions, while AES can compute similarity for pairs of samples.

3. Experiments

3.1. Datasets

Amazon Product Reviews The Amazon product review dataset (He and McAuley, 2016; McAuley et al., 2015) is a collection of domain-specific sub-datasets. Each sub-dataset is of varying size, and contains user product reviews from Amazon.com. The `summary` attribute is used as reference headline (Ma et al., 2018). We train the language models on the 'patio, lawn and garden' dataset.

CNN/Daily Mail The CNN/Daily Mail dataset (Hermann et al., 2015) is composed of news stories collected from *cnn.com* and *dailymail.co.uk*. While explicit headlines are not provided, each story has multiple 'highlights' – key takeaways from the story. For our experiments, we use the first highlight as the ground-truth/reference headline for that story. The dataset, though originally created for the question/answering task, was adapted for summarization by Nallapati et al. (2016).

Gigaword Another standard dataset used in text summarization is the annotated Gigaword coprus (Napoles et al., 2012). The dataset contains 10 million articles, each having a corresponding headline. The headline has been previously used as a summary (Rush et al., 2015). The length of each story in Gigaword is much shorter compared to other datasets used in our experiments.

Newsroom Newsroom (Grusky et al., 2018) is a recently released news-centric dataset specifically aimed at text summarization tasks. It is composed of English news-related articles produced by 38 notable publications. The authors claim that the dataset captures a variety of human summarization styles, making it amenable to abstractive, extractive and mixed summarization strategies.

For our experiments, we take approximately 90,000 story/headline pairs from CNN/DailyMail, Newsroom and Gigaword each. These are then split into train/test sets. As Amazon's 'patio, lawn and garden' is much smaller than the rest, we use the whole dataset for our experiments. Each dataset is split into 90% training data and 10% test data, the latter of which is used to generate headlines.

3.2. Training

For training, we follow a modified version of the approach introduced by Radford et al. (2019). The authors evaluate the quality of summarization using GPT-2 without further fine-tuning on CNN/Daily Mail. We improve the quality of generated headlines, by fine-tuning GPT-2 and ULMFiT on the review/story text and reference headlines. The training data has the following format:

```
Review/Story Text + [TL;DR:] + Headline +
[End]
```

The model learns how a reference headline should look like given the full review. The `[TL;DR:]` token signals to the model the end of the review and start of the headline. During headline generation, the model is then given the input:

```
Review/Story Text + [TL;DR:]
```

Both ULMFiT and GPT-2 are conditioned and fine-tuned for Amazon reviews, though only GPT-2 is subsequently also trained for CNN/Daily Mail, Gigaword, and Newsroom (discussed further in section 4.1.). The datasets are trained using the same scheme as above. For each dataset, GPT-2 is trained for 5000 counters with *learning rate = 1e-4*, and headlines are generated with *top-k = 40* and *temperature = 1.0*. The randomness of text generation can be

Dataset	Metric	Sample size	Gen-Story		Ref-Story		Gen-Ref	
			r	p-value	r	p-value	r	p-value
Amazon	ROUGE-1	20	0.1199	0.6148	0.2100	0.3743	0.602	0.00498
	ROUGE-2	20	-0.0746	0.75446	0.2417	0.30459	0.5423	0.0135
	ROUGE-L	20	0.0843	0.72379	0.1959	0.40773	0.575	0.008
	AES	20	-0.0330	0.89112	0.1260	0.59791	**0.6540**	**0.00176**
CNN/Daily Mail	ROUGE-1	20	0.3785	0.09985	-0.1565	0.51001	0.4875	0.02923
	ROUGE-2	20	0.5019	0.02414	-0.3061	0.18933	0.4681	0.0374
	ROUGE-L	20	0.4346	0.05549	-0.1802	0.44702	0.4436	0.05007
	AES	20	0.2430	0.30174	0.0000	0.99916	0.2550	0.27735
Newsroom	ROUGE-1	20	-0.0465	0.84576	0.5390	0.01419	**0.7706**	**7e-05**
	ROUGE-2	20	-0.1231	0.60523	0.2858	0.22184	0.6027	0.00491
	ROUGE-L	20	0.0563	0.81366	0.4891	0.02865	**0.7560**	**0.00012**
	AES	20	0.5010	0.02449	0.4110	0.07149	**0.8240**	**1e-05**
Gigaword	ROUGE-1	20	0.5450	0.01295	0.4636	0.0395	**0.6584**	**0.0016**
	ROUGE-2	20	0.5522	0.01158	0.1640	0.48955	0.3722	0.10613
	ROUGE-L	20	0.5056	0.02296	0.4976	0.02559	0.5985	0.00531
	AES	20	0.4070	0.07509	**0.7260**	**0.00029**	0.4480	0.04759
Overall	ROUGE-1	80	0.2829	0.01099	0.2308	0.03941	**0.6749**	**6.6e-12**
	ROUGE-2	80	0.3023	0.00643	0.0797	0.48196	**0.5128**	**1.1e-06**
	ROUGE-L	80	0.2878	0.00963	0.2570	0.02136	**0.6542**	**4.69e-11**
	AES	80	0.2290	0.04128	0.2820	0.01127	**0.5810**	**2e-08**

Table 3: Summary of Pearson correlation of human judgment with automatic metrics. Bold *r* and *p-value* pairs indicate Bonferroni-adjusted statistically significant results (*p-value* less than 0.0042)

adjusted by the latter two hyperparameters, and we observe that using the aforementioned values for each strikes a sufficient balance between relevant and creative summaries in our case.

Pre-processing steps on the dataset include filtering out useless phrases, such as time, place, or author of a story and clipping each story to a maximum of five sentences. In the case of Amazon reviews, shorter headlines (less than 15 characters) are filtered out. This proved to be useful for generating more meaningful headlines as the shorter headlines are generic and not informative of the product.

3.3. Human Evaluation

We perform three manual evaluation tasks in order to compare the proposed automatic evaluation metrics to human judgments. For the first task, the person is asked to read a story text and decide if the generated headline is a reasonable summary of the respective story. The sentiment and content of a story and a correspondingly generated headline are supposed to be related, and the headline text should be understandable and not artificial. Similarly, the second task asks the person to assess the same, only this time for the reference headline instead of the generated one. The third and final task is asking the person to judge the similarity between the reference and generated headlines. The tasks are termed as 'Gen-Story', 'Ref-Story' and 'Gen-Ref' respectively. The testers are kept ignorant of whether a given headline is real or generated. Five testers are asked to score the respective similarities on a scale from 1-5, where 5 means very similar and 1 hardly similar.

All three tasks are performed for each of the four datasets, with 20 samples taken from each. This gives each tester a total of 80 samples with three tasks.

4. Results

4.1. Comparison of Language Models

We initially test which language model is more suitable to generate relevant headlines. This is done by fine-tuning both GPT-2 and ULMFiT on the 'patio, lawn and garden' dataset, generating headlines for said dataset and evaluating them using AES and human judgment. The results can be seen in table 2. GPT-2 clearly outperforms ULMFiT which is consistent with human assessment. Henceforth, we only use GPT-2 for our subsequent experiments to test the validity of AES and FED as automatic metrics.

Table 1 shows a good, average, and bad example of the capabilities of GPT-2. Note that the model still remembers that Chicago lies in the Midwest of the US from its pre-training on WebText. More examples can be found in our repository.

4.2. Metric Evaluation

In this section, we report whether AES and FED relate with human judgment in a significant manner. We also compare AES with the standard metric ROUGE.

AES We use Pearson correlation as the measure to gauge the correlation between two variables, and perform null hypothesis testing using p-values. As AES can be calculated on a per-sample basis, we have AES scores for 20 samples per dataset (80 in total) for each of the 3 tasks. The tasks are listed as 'Gen-Story', 'Ref-Story' and 'Gen-Ref' and their description can be found in section 3.3. This gives us a total of 240 AES scores (80 per task). As we have 5 human evaluators, we calculate the human average for each

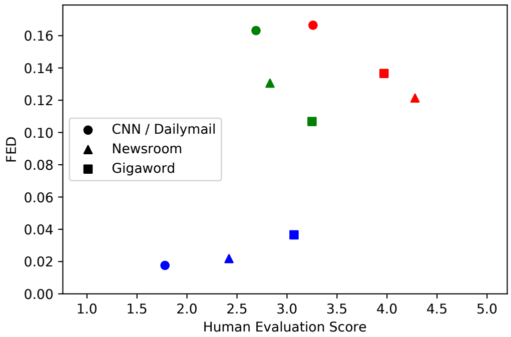

Figure 2: Scatter plot of FED vs human evaluation. The three tasks are color-coded: Green denotes ’Gen-Story’, orange denotes ’Ref-Story’ and blue denotes ’Gen-Ref.’

task per sample to match the count of 240 AES scores. Finally, the correlation between the AES scores and average human judgment is calculated for each task as well as the significance of the correlation via p-values.

Table 3 shows the results of the metric evaluation per dataset, as well as an overall assessment. ROUGE values are also listed for the sake of comparison. As can be observed, AES correlates positively with human judgment in all but two cases. Many of the positive correlations are also statistically significant.

Table 4 shows how many times a metric was the highest correlated one with human judgment. Note that this is done by looking at each task in table 3 for each dataset and noting which metric has the highest *r* value. Although per convention a p-value below 0.05 is considered statistically significant, we perform the Bonferroni adjustment (Weisstein, 2004) which corrects for the number of experiments done with a dataset. In our case, 12 experiments are done on each dataset. Dividing 0.05 by 12 yields a Bonferroni-corrected statistical significance threshold of 0.0042. The results are reported in table 4. A metric may have a statistically significant correlation with human evaluation while never having the highest *r* value, as in the case of ROUGE-L. The table provides a clear picture of AES when compared to ROUGE; AES is highest correlated with human perception more frequently than any ROUGE metric individually, and the correlations are statistically significant more often than any ROUGE metric.

FED In contrast to AES, FED can only be calculated on a per-corpus basis, rather than a per-sample basis as stated in section 2.2.2. The efficacy of calculating *r* between human judgment and FED is thus diminished due to low sample size (we have 12 FED values in total as a result of all experiments). However, the relation between human perception and FED can still be demonstrated, albeit in a less robust manner compared to AES, by plotting the average human scores against FED values. This can be seen in figure 2. Human comparison with Amazon’s ’patio, lawn and garden’ is omitted from the figure as the values were considered outliers (more than 10x those of other datasets). We hypothesize that this is due to the vast number of uninformative headlines in that corpus.

A clear trend in the first two tasks (colored green and orange) is that an increase in human scores results in lower FED, as hypothesized. The last task shows no strong relation.CNN/Daily Mail have high FED values in task 2 and task 3 even though they follow the expected trend of negative relation with human judgment. We speculate that this is due to the fact that the dataset does not contain headlines for each story, rather containing multiple ’highlights’ which emphasize key points of the story. As such, any single highlight may be unable to capture the crux of the story. FED’s sensitivity to sentence-length is also demonstrable as the values for task 3 (blue) are visibly smaller than the other two tasks that involve the story/review text.

Metric	# best correlations	Bonferroni corrected
ROUGE-1	4	3
ROUGE-2	5	1
ROUGE-L	0	2
AES	6	4

Table 4: Table showing how many times each metric was the highest correlated one with human judgment. Additionally, we can see how many of the correlations were statistically significant after applying the Bonferroni correction.

5. Conclusion

In this paper, we fine-tune and condition language models to generate abstractive summaries in the form of headlines. Qualitatively, many generated headlines appear to be valid for the given text. To further evaluate the headlines, we rely on the recently published FED (Semeniuta et al., 2018; Conneau et al., 2017) as well as on AES, which is our proposed metric. Experimentally, we show that AES corresponds to human perception and performs mostly better than the traditional ROUGE metric, whereas FED does not always relate to human perception not least due to its sensitivity to text length.

All evaluated metrics for abstractive summarization are merely based on the pre-trained Universal Sentence Encoder (Cer et al., 2018). However, recently many other textual embeddings have been published which might be a better choice with respect to computational efficiency, e.g. smooth inverse frequency (Arora et al., 2017), accuracy, e.g. BERT (Devlin et al., 2018), or stability with respect to length of text, e.g. doc2vec (Le and Mikolov, 2014). In particular the latter could potentially lead to the development of metrics which would be more suitable for abstractive summarization tasks other than headline generation.

6. Bibliographical References

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. *ICLR*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

d'Autume, C. d. M., Rosca, M., Rae, J., and Mohamed, S. (2019). Training language gans from scratch. *arXiv preprint arXiv:1905.09922*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fréchet, M. (1957). Sur la distance de deux lois de probabilité. *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES*, 244(6):689–692.

Graham, Y. (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.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637.

Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Hsu, W.-T., Lin, C.-K., Lee, M.-Y., Min, K., Tang, J., and Sun, M. (2018). A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

Iyyer, M., Manjunatha, V., Boyd-Graber, J. L., and Daumé, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In *ACL*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Lin, C.-Y. and Hovy, E. (2003). The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 73–80. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moratanch, N. and Chitrakala, S. (2016). A survey on abstractive text summarization. In *2016 International Conference on Circuit, power and computing technologies (ICCPCT)*, pages 1–7. IEEE.

Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. (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.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *CoRR*, abs/1509.00685.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Semeniuta, S., Severyn, A., and Gelly, S. (2018). On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*.

Smith, L. N. (2015). No more pesky learning rate guessing games. *CoRR*, abs/1506.01186.

Takase, S., Suzuki, J., Okazaki, N., Hirao, T., and Nagata, M. (2016). Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1054–1059.

Weisstein, E. W. (2004). Bonferroni correction.

7. Language Resource References

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Ma, S., Sun, X., Lin, J., and Ren, X. (2018). A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *arXiv preprint arXiv:1805.01089*.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 43–52, New York, NY, USA. ACM.

Napoles, C., Gormley, M., and Van Durme, B. (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.