# Spotify at TREC 2020:
# Genre-Aware Abstractive Podcast Summarization

**Rezvaneh Rezapour***
University of Illinois at Urbana-Champaign
Urbana, IL, USA
rezapou2@illinois.edu

**Sravana Reddy**
Spotify
Boston, MA, USA
sravana@spotify.com

**Ann Clifton**
Spotify
New York, NY, USA
aclifton@spotify.com

**Rosie Jones**
Spotify
Boston, MA, USA
rjones@spotify.com

## Abstract

This paper contains the description of our submissions to the summarization task of the Podcast Track in TREC (the Text REtrieval Conference) 2020. The goal of this challenge was to generate short, informative summaries that contain the key information present in a podcast episode using automatically generated transcripts of the podcast audio. Since podcasts vary with respect to their genre, topic, and granularity of information, we propose two summarization models that explicitly take genre and named entities into consideration in order to generate summaries appropriate to the style of the podcasts. Our models are abstractive, and supervised using creator-provided descriptions as ground truth summaries. The results of the submitted summaries show that our best model achieves an aggregate quality score of 1.58 in comparison to the creator descriptions and a baseline abstractive system which both score 1.49 (an improvement of 9%) as assessed by human evaluators.

## 1 Introduction

Advancements in state-of-the-art sequence to sequence models and transformer architectures (Vaswani et al., 2017; Dai et al., 2019; Nallapati et al., 2016; Ott et al., 2019; Lewis et al., 2020), and the availability of large-scale datasets have led to rapid progress in generating near-human-quality coherent summaries with abstractive summarization. The majority of effort in this domain has been focused on summarizing news datasets such as CNN/Daily Mail (Hermann et al., 2015) and XSum (Narayan et al., 2018). Podcasts are much more diverse than news articles in the level

---

* Work done while at Spotify

of information as well as their structure, which varies by theme, genre, level of formality, and target audience, motivating the TREC 2020 podcast summarization task (Jones et al., 2020), where the objective is to design models to produce coherent text summaries of the main information in podcast episodes.

While no ground truth summaries were provided for the TREC 2020 task, the episode descriptions written by the podcast creators serve as proxies for summaries, and are used for training our supervised models. We make some observations about the creator descriptions towards understanding what goes into a 'good' podcast summary. Our first observation is that the descriptions vary stylistically by genre. For instance, the descriptions of sports podcasts tend to list a series of small events and interviews, usually anchored on the names of players, coaches, and teams. In contrast, descriptions of true crime podcasts tend to contain suspenseful hooks, and avoid giving away much of the plot (Table 1). Some podcasts include the names of hosts or guests in their descriptions, while others only talk about the main theme of the episode.

By definition, a good summary should be concise, preserve the most salient information, and be factually correct and faithful to the given document (Nenkova and McKeown, 2011; Maynez et al., 2020). In addition, we believe that since users' expectations of summaries are shaped by the creator descriptions and the style and tone of the content, it is important to generate summaries that are appropriate to the style of the podcast.

In this work, we present two types of summarization models that take these observations into consideration. The goal of the first model, which we shorthand as 'category-aware', is to generate

| | |
|---|---|
| **Sports** | In EP 4 *Jon Rothstein* is joined by Arizona State Head Coach *Bobby Hurley*, Creighton Guard *Mitchell Ballock*, and Merrimack Head Coach *Joey Gallo* checks in on the Hustle Mania Hotline. This is *March* and *Jon* has a special message for all of you heading into the NCAA tournament. |
| **True Crime** | Sometimes, *it takes years to connect a killers crimes*. On March 6th 1959 *a man* was born who would, eventually, be tried for the murder of a single woman. I*t would take years for police to connect him to 4 other murders*. Years and a clever investigator who got the DNA he needed. |
| **Comedy** | This weeks episode is a little bit of a throwback.. We have *one of our close high school friends Zak on the show*, and he is just a *blast*! We discuss relationships, our favourite movies and *our experience's with high school*! This episode is pretty chill, so you better have a drink for this one! |

Table 1: Examples of creator descriptions for a different genres of podcast episodes, highlighting the prevalence of named entities in sports, suspense in true crime, and colloquial language in comedy.

summaries whose style is appropriate for the genre or category of the specific podcast. The second model aims to maximize the presence of named entities in the summaries.

## 2 Data Filtering

The Spotify Podcast Dataset consists of 105,360 episodes with transcripts and creator descriptions (Clifton et al., 2020), and is provided as a training dataset for the summarization task. We applied a set of heuristic filters on this data with respect to the creator descriptions.

- Removed episodes with unusually short (fewer than 10 characters) or long (more than 1300 characters) creator descriptions.

- Removed episodes whose descriptions were highly similar to the descriptions of other episodes in the same show or to the show description. Similarity was measured by taking the TF-IDF representation of a description and comparing it to others using cosine similarity.

- Removed ads and promotional sentences in the descriptions. We annotated a set of 1000 episodes for such material, and trained a classifier to detect these sentences; more details are described in Reddy et al. (2021).

The dataset after these filters consists of 90055 episodes. We held out 1000 random episodes for test and validation sets for development, and used 88055 episodes for training our models. Our final submissions were made on the task's test set of 1027 episodes disjoint from the training data.

## 3 Abstractive Model Framework

Our summarization models are built using BART (Lewis et al., 2020), which is a denoising autoencoder for pretraining sequence to sequence models. To generate summaries, we started from a model pretrained on the CNN/Daily Mail news summarization dataset[1] and then fine-tuned it on our podcast transcript dataset with respect to the two proposed models (described below in §4).

We used the implementation of BART in the Huggingface Transformers library (Wolf et al., 2020) with the default hyperparameters. We set the batch size to 1, and made use of 4 GPUs for training. The best model with the highest ROUGE-L score on the validation set was saved. The maximum sequence limit we set for the input podcast transcript was 1024 tokens (that is, any material beyond the first 1024 tokens was ignored). We constrained the model to generate summaries with at least 50 and at most 250 tokens.

This setup is the same as the *bartpodcasts* model described in the task overview (Jones et al., 2020), and is one of our baselines for comparison. We also compare our models to the *bartcnn* model baseline, which is the out of the box BART model trained on the CNN/Daily Mail corpus.

## 4 Models

### 4.1 Category-Aware Summarization

The idea of building a category-aware summarization model is motivated by the hypothesis that selecting what is important to a summary depends on the topical category of the podcast. This hypothesis is based upon observations of the creator descriptions in the dataset. For example, descriptions of 'Sports' podcast episodes tend to contain names of players and matches, 'True Crime' podcasts incorporate a hook and suspense, and 'Comedy' podcasts are stylistically informal (Table 1).

While we expect some of these patterns to be naturally learned by a supervised model trained on the large corpus, we nudged the model to generate

---

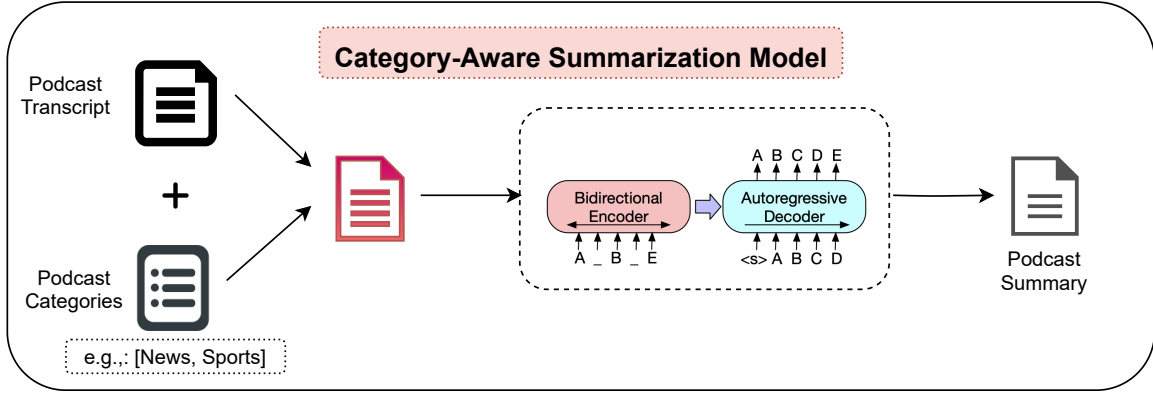[1] https://huggingface.co/facebook/bart-large-cnn

Figure 1: Sketch of the Category Aware Summarization Model

stylistically appropriate summaries by explicitly conditioning the summaries on the podcast genre. Some previous work approaches this problem by concatenating a vector corresponding to an embedding of the conditioning parameters to the inputs in an RNN (Ficler and Goldberg, 2017). More recent work simply concatenates a token corresponding to the parameter to the text input in sequence-to-sequence transformer models (Raffel et al., 2020). We experimented with a few ways of encoding the category labels in the summarization model, and settled upon prepending the labels as special tokens to the transcript as the input during training and inference, as shown in Figure 1. For the episodes with multiple category labels, we concatenated the labels and included them all as distinct special tokens, concatenated in a fixed (lexicographic) order.

The podcast categories are not given in the metadata files distributed with the podcast dataset. However, they can be derived from the RSS headers associated with each podcast under the *itunes:category* field. In our case, we leveraged the labels assigned in the Spotify catalog. The category labels in the catalog are mostly the same as the labels in the RSS header, with some minor changes. The taxonomy of iTunes categories changes over time, leading to different labels for semantically similar categories. We chose to heuristically collapse similar categories together, such as 'Sports' with 'Sports & Recreation', giving a set of 22 categories (Table 2).

After prepending the labels as special tokens to the transcript, we fine-tuned the BART baseline. Since training BART was computationally expensive, we only trained the model for a maximum of 2 epochs and generated abstractive summaries using 1 (*category-aware-1epoch*) and 2 (*category-*

| Categories | Counts |
|---|---|
| Arts | 81 |
| Business | 97 |
| Comedy | 136 |
| Education | 118 |
| Fiction | 5 |
| Games | 7 |
| Government | 3 |
| Health | 94 |
| History | 16 |
| Kids & Family | 28 |
| Leisure | 45 |
| Lifestyle & Health | 42 |
| Music | 33 |
| News | 15 |
| Religion & Spirituality | 92 |
| Science | 11 |
| Society & Culture | 96 |
| Sports | 137 |
| Stories | 37 |
| TV & Film | 41 |
| Technology | 17 |
| True Crime | 28 |

Table 2: Set of collapsed category labels and their distribution on the task test set. Some episodes are associated with multiple labels.

*aware-2epochs*) epochs, separately.

## 4.2 Named Entity Biased Model

Named entities are information-dense, and many podcast topics tend to center around people and places, making them appropriate for inclusion in summaries. We furthermore (1) observed that named entity mentions are prevalent in creator descriptions, and (2) surveyed a group of podcast listeners about what they look for in a summary, where the presence of names and locations was highlighted. Therefore, our second summarization algorithm aims to bias the model to maximize named entities in summaries.

Our model takes a two-step approach by extracting a portion of the transcript that is both highly

relevant to the episode and tends to contain salient named entities. The extracted portion of the transcripts is the input for both training and inference. This approach is similar to the previous approaches that implement an extractive summarizer followed by an abstractive model (Tan et al., 2017).

Our proposed extractive summarizer is similar to TextRank (Mihalcea and Tarau, 2004) and consists of the following steps:

- We first identified named entities for the entire transcript of each episode by using a BERT token classification model trained[2] on the CoNLL-03 NER data.

- We divided each episode into segments corresponding to about 60 seconds of audio.

- Every segment was represented in its original form $s$, and as a list of its named entity tokens $t$. All pairwise similarities between segments were computed as the proportion of word overlap between them. We computed both $Sim_s(i, j)$, the similarity between the original forms of the segments $i$ and $j$, and $Sim_t(i, j)$, the similarity between the list of named entity tokens in the segments.

- Degree centralities for each segment $C_s(i) = \sum_j Sim_s(i, j)$ and $C_t(i) = \sum_j Sim_t(i, j)$ were computed. The top 7 segments by $C_s(i)$ and the top 3 non-overlapping segments by $C_t(i)$ were extracted and presented in position order as the extractive summary. We chose this heuristic to encourage the extracted segments to be semantically relevant and contain named entities.

The BART baseline was then fine-tuned for 3 epochs on extracted segments as input (rather than the first 1024 tokens of the transcript). We name this the *coarse2fine* model.

| | ROUGE-L | | |
|---|---|---|---|
| | **Precision** | **Recall** | **F1** |
| *bartcnn* | 8.49 | **27.19** | 11.3 |
| *bartpodcasts* | 20.78 | 21.01 | 16.64 |
| ***category-aware-1epoch*** | 22.70 | 20.82 | 17.62 |
| ***category-aware-2epochs*** | **25.75** | 19.86 | **18.42** |
| *coarse2fine* | 15.76 | 18.75 | 13.59 |

Table 3: ROUGE scores against all of the creator descriptions in the test set.

| | ROUGE-L | | |
|---|---|---|---|
| | **Precision** | **Recall** | **F1** |
| *bartcnn* | 10.87 | **29.85** | 14.64 |
| *bartpodcasts* | 27.89 | 25.51 | **22.15** |
| ***category-aware-1epoch*** | 27.49 | 22.45 | 20.54 |
| ***category-aware-2epochs*** | **30.63** | 23.23 | 22.02 |
| *coarse2fine* | 22.85 | 20.57 | 17.12 |

Table 4: ROUGE scores aggregated over only those 71 creator descriptions that were rated good or excellent.

| | E | G | F | B | score |
|---|---|---|---|---|---|
| Creator Description | 15.64 | 24.02 | 34.08 | 26.26 | 1.45 |
| Cleaned Creator Description | 18.44 | 21.23 | 32.96 | 27.37 | 1.49 |
| *bartcnn* | 5.59 | 13.97 | 49.26 | 31.28 | 0.99 |
| *bartpodcasts* | 13.97 | 27.93 | 37.43 | 20.67 | 1.49 |
| ***category-aware-1epoch*** | 14.53 | 27.37 | 38.55 | 19.55 | 1.51 |
| ***category-aware-2epochs*** | 17.88 | 21.79 | 43.02 | 17.32 | 1.58 |
| *coarse2fine* | 10.06 | 21.79 | 29.61 | 38.55 | 1.30 |

Table 5: Distribution of EGFB annotations over 179 judged episodes (percentages), and the aggregate quality score, computed as a weighted average with E=4, G=2, F=1, and B=0.

## 5 Results

The TREC task uses two evaluation methods to analyze performance: (1) ROUGE-L scores to compare the generated summaries against the creator descriptions as the reference, and (2) human evaluations facilitated by the National Institute of Standards and Technology (NIST) to asses the quality of the generated summaries with respect to the episode content.

Table 3 shows the results of our models against the *bartcnn* and *bartpodcasts* baselines on the whole test set. Compared to the baselines, both *category-aware* models achieved higher ROUGE precision and F1-scores. While *bartcnn* has the highest recall (27.19%), our *category-aware-2epochs* summarizer outperformed other models achieving 27.75% precision and 18.42% F1-score.

Since creator descriptions are of varying quality, we performed another evaluation, selecting only those episodes whose creator descriptions were labeled as 'Good' or 'Excellent' by the evaluators (Table 4). The *category-aware-2epochs* model still achieves the highest ROUGE precision, and has comparable F1 with the *bartpodcasts* baseline.

High precision across our models shows that prepending the categories to the transcripts resulted in more *relevant* information in the generated summaries, with tradeoffs against recall. Tables 3 and 4 show that the *coarse2fine* model only improved over *bartcnn*. We observe that the summaries do consist of more named entities compared to the other models (Table 6) as intended. We observe

that the discontinuities in the extracted segments result in more incoherent abstractive summaries than the models that consume a contiguous block of text, and sometimes contain 'hallucinated' content.

Looking into the manual assessment of the summaries (Table 5), we see that the *category-aware-2epochs* model generated more summaries labeled as 'Excellent' and fewer summaries labeled as 'Bad' compared to the other models and is comparable in quality to the creator descriptions.

As an illustrative example, we present the summaries of the models as well as the creator description of one selected episode of a sports podcast in Table 6. The creator description contains different named entities (in **bold**) as well as social media information and links (in blue). The *bartcnn* summary is verbose and long and captured the majority of the information from the beginning of the transcript, following the structure of the news summarizer. *bartpodcasts* and *category-aware-1epoch*, on the other hand, are similar to the creator description. The *category-aware-2epochs* is the shortest compared to the other models, however, it includes the main theme and the majority of the necessary information. Finally, the *coarse2fine* summary does not include the names of the host or guest but it is the only one that contains additional named entities, similar to the creator description.

NIST judges also assessed the quality of the summaries by answering eight yes/no questions about podcast summary attributes. A good summary will receive 'yes' for all questions except Q6 ("Does the summary contain redundant information?"). With respect to Q6, Q7, and Q8, our *category-aware-2epoch* model generates the most coherent summaries with the least amount of redundant information compared to the baselines (Table 7) suggesting that the model may be learning aspects of style.

Overall, the automatic and manual evaluations of our models show that our *category-aware* models outperformed the baseline resulting in high quality summaries. The results of the submitted summaries show that our best model achieves an aggregate quality score (as defined by the task) of 1.58 in comparison to the creator descriptions and a baseline abstractive system which both score 1.49 (an improvement of 9%) as assessed by human evaluators (Table 5).

## 6    Conclusion

In this paper, we showcased two abstractive summarization models, one that is informed by the categories of the podcasts, and a hybrid model that uses an extractive step biased towards named entities. Experimental results show that our category-aware models generate better summaries than the baselines, but the model intended to bias summaries to named entity mentions underperforms. In future, we hope to investigate new and robust ways of biasing the summarizer to generate named entities in the summaries.

## References

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2020. Overview of the TREC 2020 Podcasts Track. In *The 29th Text Retrieval Conference (TREC 2020) notebook*. NIST.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

| Creator Description | **Mike Rankin** welcomes Sox in the Basement host **Chris Lanuti** to talk optimism on the South Side. The guys react to **Yoan Moncada**'s recent extension and later explore the potential for **Dylan Cease**, **Aaron Bummer** and well as **Andrew Vaughn** in 2020. You can follow FutureSox (@FutureSox), Chris Lanuti (@ChrisLanuti) and Sox In The Basement (@SoxInTheBasement) on Twitter. |
|---|---|
| **Generated Summaries** | |
| *bartcnn* | Hi. Hello and welcome to another episode of the Future Socks podcast. My name is **Mike Rankin**. I will be your host this time around and boy. What's going on? I appreciate you jumping in with us to talk some White Sox baseball on the future Sox podcast. You talked about socks in the basement for a couple of years now take me back to the Art, and what was the inspiration and then the growth across those two seasons? Well, the funny thing is that I had been doing podcasting for like a decade just as a hobby I used to do morning radio, but I never did anything in Chicago. I was in Champaign, Illinois and Reno, Nevada and Wheeling, West Virginia Pittsburgh and Southern California and Bakersfield, which is the armpit of America. It's like a hundred twenty two degrees in the shade and I'm from Chicago. And so when I came back here and you know started doing Thing that was out of radio. I still had the bug and I did this show called the broadcast basement and it was just a bunch of my friends used to be in radian. We'd mess around and somewhere along the line a couple years ago. |
| *bartpodcasts* | Socks in the Basement" host **Chris Ludacris** joins **Mike Rankin** of the Future Sox Podcast to talk all things White Sox baseball, including the recent success of the team and what it means for the future of the franchise. Chris is the host of the Socks In The Basement podcast, which is a weekly podcast covering the Chicago White Sox. You can follow them on Twitter @socksinthebasepod and check out the show at socksinthesound.com. You can also follow the show on Twitter at @SocksInTheBasetPod and on Instagram at SocksinTheBasepod.comsocks_in_the_baset. The show is brought to you by the Chicago Sun-Times and Wrigley Field. |
| *category-aware-1epoch* | On this episode of the Future Sox Podcast, **Chris Ludacris** from Socks in the Basement joins **Mike Rankin** to talk about the White Sox's 2019 season and what we expect from the Sox in 2020. We also talk about how the Sox are getting better on the field and off the field, and how we expect the Sox to get better in 2020 as well. We hope you enjoy this episode, and if you do, please let us know what you thought of the show and what you would like to see in 2020 by leaving us a review on your favorite podcast platform. We love to hear your thoughts and feedback on the show, so feel free to reach out to us on Twitter at @SocksInTheBasementPod and we'd love to have you on the podcast in the future. |
| *category-aware-2epochs* | **Chris Ludacris** of Socks in the Basement joins the show to talk about the White Sox offseason and the future of the team as well as his podcast Socks In The Basement and how it has grown to where it is today. |
| *coarse2fine* | Socks in the Basement"" is a weekly podcast covering the Chicago White Sox baseball team through the eyes of Sox fans. Hosts are @BBWhiteSox, @SoxOnTheBus, and @BBBlackSox. This week we discuss the White Sox's bullpen, **Jose Abreu**, **Jose Ramirez**, **Dylan Cease**, and much more. You can follow us on Twitter @SocksInTheBasementPod and check out the extra content about the show at SoxOnTheBaset.com. |

Table 6: Example summaries generated by various models for a sports podcast episode (`https://open.spotify.com/episode/5KmWK8Qh5lsWIb2sUuJp4r`).

| Attribute | *bartcnn* | *bartpodcasts* | *category-aware-1epoch* | *category-aware-2epochs* | *coarse2fine* |
|---|---|---|---|---|---|
| Q1: Does the summary include names of the main people (hosts, guests, characters) involved or mentioned in the podcast? | **72.6** | 51.4 | 64.2 | 55.3 | 35.2 |
| Q2: Does the summary give any additional information about the people mentioned (such as their job titles, biographies, personal background, etc)? | **48.0** | 34.6 | 40.2 | 40.8 | 22.9 |
| Q3: Does the summary include the main topic(s) of the podcast? | 69.3 | **78.2** | 76.5 | 77.7 | 60.3 |
| Q4: Does the summary tell you anything about the format of the podcast; e.g. whether it's an interview, whether it's a chat between friends, a monologue, etc? | 58.1 | 49.2 | **59.2** | 57.0 | 41.3 |
| Q5: Does the summary give you more context on the title of the podcast? | **67.6** | 65.4 | 65.4 | 65.9 | 57 |
| Q6: Does the summary contain redundant information? | 25.7 | 14.0 | 12.8 | **6.7** | 10.1 |
| Q7: Is the summary written in good English? | 34.6 | 86.6 | 86 | **88.0** | 79.9 |
| Q8: Are the start and end of the summary good sentence and paragraph start and end points? | 22.3 | 62 | 69.8 | **70.4** | 55.3 |

Table 7: Average percentage of 'Yes' judgments by human evaluators for each of the summary attributes.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, 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, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. *Automatic summarization*. Now Publishers Inc.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting extraneous content in podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4109–4115.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.