TREC 2020 Podcasts Track Overview

Rosie Jones[1], Ben Carterette[1], Ann Clifton[1], Maria Eskevich[2],
Gareth J. F. Jones[3], Jussi Karlgren[1],
Aasish Pappu[1], Sravana Reddy[1], Yongze Yu[1]

[1] Spotify
[2] CLARIN ERIC
[3] Dublin City University

Abstract The Podcast Track is new at the Text Retrieval Conference (TREC) in 2020. The podcast track was designed to encourage research into podcasts in the information retrieval and NLP research communities. The track consisted of two shared tasks: segment retrieval and summarization, both based on a dataset of over 100,000 podcast episodes (metadata, audio, and automatic transcripts) which was released concurrently with the track. The track generated considerable interest, attracted hundreds of new registrations to TREC and fifteen teams, mostly disjoint between search and summarization, made final submissions for assessment. Deep learning was the dominant experimental approach for both search experiments and summarization. This paper gives an overview of the tasks and the results of the participants' experiments. The track will return to TREC 2021 with the same two tasks, incorporating slight modifications in response to participant feedback.

1 Introduction

Podcasts are a growing medium of recorded spoken audio. They are more diverse in style, content, format, and production type than previously studied speech formats, such as broadcast news (Garofolo et al., 2000) or meeting transcripts (Renals et al., 2008), and they encompass many more genres than typically studied in video research (Smeaton et al., 2006). They come in many different formats and levels of formality – news journalism or conversational chat, fiction or non-fiction. Podcasts have a sharply growing share of listening consumption (Edison Research, 2020) and yet have been relatively understudied. The medium shows great potential to become a rich domain for research in information access and speech and language technologies (among other fields), with many potential opportunities to improve user engagement and consumption of podcast content. The TREC Podcast Track which was launched in 2020 is intended to facilitate research in language technologies applied to podcasts.

1.1 Data

The data distributed by the track organisers consisted of just over 100,000 episodes of English-language podcasts. Each episode comes with full audio, a transcript which was automatically generated using Google's Speech-to-Text API as of early 2020, and a description and metadata provided by the podcast creator, along with the RSS feed content for the show. The data set is described in greater detail by Clifton et al. (2020); an example is given in Figure 1.

Statistic Name	Value
Email list sign-ups	285
In TREC slack channel #podcasts 2020	194
TREC podcasts registrations	213
Signed data sharing agreement	77
Downloaded transcripts	64
Downloaded audio	18
Participated in Search	7
Participated in Summarization	8
Participated in Both	2

Table 1: Participation statistics

1.2 Participation

The Podcast Track attracted a great deal of attention with more than 200 registrations to participate. Most registrants did not submit experiments for assessment. After the submission deadline had passed, registrants were sent a questionnaire to establish what they found to be the biggest challenge when working on their experiment and their submission, and if they did not submit a result, what the most important challenge they found to stand in the way of submission. Participants were also asked to suggest how participation might be made easier for the coming year. The response rate was on the low side (10 responses) and the collated results indicate that the size of the data overwhelmed some participants. Suggestions for the coming year included organising a task for a subset of the data to enable new entrants to familiarise themselves with the problem space.

1.3 Tasks

In 2020 the Podcast Track offered two tasks: (1) retrieval of fixed two-minute segments and (2) summarization of episodes. Both tasks were possible to complete on the automatic transcripts of episodes, rather than the audio data. The full audio data was provided, and teams were free to use it for their tasks (though only one team did do so, using the audio to improve the automatic transcription quality). The segment retrieval and summarization submissions were entirely based on textual input for all submitted experiments.

2 Previous Work

While there has been relatively little published work exploring information access technologies for podcasts, there is longstanding interest in spoken content retrieval in a range of other settings involving spoken content.

2.1 Spoken Document Retrieval

The best known of work in spoken document retrieval is the TREC Spoken Document Retrieval Track which ran at TREC from 1997-2000 (Garofolo et al., 2000). The track focused on examining spoken document retrieval for broadcast news from radio and television sources of increasing size and complexity with each edition of the task. Task participants were provided with baseline transcripts of the spoken created using a then state-of-the-art automatic speech recognition (ASR) system and accurate or near-accurate manual transcripts of the content. The track began by using documents created by manually segmenting the news broadcasts into stories, but latterly began to explore automated identification of start points within unsegmented news broadcasts. The key findings were that for broadcasts, similar retrieval effectiveness could be achieved for errorful automatic speech recognition transcripts as for manual transcripts, through the appropriate use of external resources such as large contemporaneous news text archives.

A very different spoken retrieval task ran at the CLEF conference in the years 2005-2007 as the Cross-Language Speech Retrieval (CL-SR) task (Pecina et al., 2008). This focused on retrieval from a large archive of oral history — spontaneous conversations in the form of personal testimonies. Participants were provided with automatic speech recognition (ASR) transcripts of the spoken content, with a diverse set of associated metadata, manually and automatically assigned controlled vocabulary descriptors for concepts presented in each oral testimony, dates and locations associated with the content discussed, manually assigned person names, and expert hand-written segment summaries of the events discussed, together with a set of carefully designed search topics. The main task was to identify starting points for cohesive stories within the each conversational testimony interview where the ground-truth story boundaries were manually assigned by domain experts. The main findings of this task were that accurate automated location of topic start points is challenging, and that, importantly, conversations of this type frequently fail to include mention of important entities within the dialogue. This means that search queries which include these entities often fail to match well with relevant content. This contrasts with search of broadcast news where such entities are mentioned very frequently to enable listeners to news updates can easily understand the events being described. Retrieval effectiveness was greatly improved by judicious use of the provided manual metadata, but it was recognised that such metadata will not be available for many spoken content archives.

Another spoken content retrieval task was offered at NTCIR from 2010-2016. This focused on search of Japanese language lectures and technical presentations. The first phase of the task focused only the retrieval of spoken content (Akiba et al., 2016) while the second phase included the additional complexity of spoken queries (Akiba et al., 2011). As well as issues for automated transcription relating of the unstructured informal nature of the spoken delivery of this content, transcription of this content introduced challenges of transcription of specialised domain specific vocabulary items. Participants were provided with a set of search topics with a requirement to locate relevant content within the transcripts. A unique feature of this dataset was the very detailed fine-granularity labelling of relevant content for each search query within the transcripts. This meant that it was possible to do very detailed analysis of the ability of search methods to identify relevant content, including the relationship between search behaviour and the accuracy of the transcription of the query search terms within the transcripts.

A further study of spoken content search was the Rich Speech Retrieval and Search and Hyperlinking tasks at Mediaeval from 2011-2015 (Larson et al., 2011; Eskevich et al., 2012; 2015). The primary search focus of this task was the identification of "jump-in" points in multimedia content based on the spoken soundtrack. In different years the task focused on different multimedia content collections. Initially the Blip10000 collection of crawled content from the `blip.tv`[1] online platform of semi-professional user generated (SPUG) content (Schmiedeke et al., 2013; Eskevich et al., 2012) and later a collection of diverse broadcast television content provided by the BBC (Eskevich et al., 2015). Participants were provided with state-of-the-art ASR transcripts of the content archives and carefully developed search queries. Tasks included known-item and ad hoc search, with relevance assessment using crowdsourcing methods. As well as confirming earlier findings in terms of automated location of useful jump-in points, there was significant focus in these tasks on how submissions should be comparatively evaluated. In particular, the trade-off between ranking of retrieved items containing relevant content and the accuracy of the identification jump-in points in retrieved items.

As well as these benchmark tasks, another relevant study in spoken content retrieval using the AMI corpus (Renals et al., 2008) is reported in Eskevich and Jones (2014) which gives a detailed examination of the differences in the ranking of retrieved items between manual and automated transcripts arising from ASR errors. A more complete overview of research in spoken content retrieval from its beginnings in the early 1990s to today can be found in Jones (2019). While none of this existing work focuses on podcast search, the various content archives used raise many of the same issues that can be observed in podcasts in terms of content diversity, use of domain specific vocabularies, and probable issues relating to absence of entity mentions in conversational podcasts.

2.2 Summarization

While there is a great deal of work on summarizing text in the news domain (eg Mihalcea and Tarau (2004)), there is much less existing work on summarization of spoken content. One study relevant to the Podcast Track is that of Spina et al. (2017). This work focuses on the creation of query biased

[1] `https://en.wikipedia.org/wiki/Blip_(website)`

audio summaries of podcasts. A crowdsourced experiment demonstrated that highly noisy automatically generated transcripts of spoken documents are effective sources of document summaries to support users in making relevance judgements for a query. Particularly notable was the finding that summaries generated using ASR transcripts were comparable in terms of usability to summaries generated using error-free manual transcripts.

2.3 Podcast Information Access

Besser et al. (2008) argues that the underlying goals of podcast search may be similar to those for blog search, as podcast can be viewed as audio blogs. In Tsagkias et al. (2010), the general appeal of podcast feeds/shows is predicted from various features. The authors identify as important factors of whether a user subscribes to a podcast feed: whether the feed has a logo, length of the description, keyword count, episode length, author count, and feed period.

Yang et al. (2019) showed they could use acoustic features to predict seriousness and energy of podcasts, as well as popularity. Acoustic features take advantage of a unique aspect of podcasts, and can be used as part of a multimodal approach to podcast information access, which we hope to see more of in the track in future years.

3 Segment Retrieval Task

3.1 Definition

The retrieval task was defined as the problem of finding relevant segments from the episodes for a set of search queries which were provided in traditional TREC topic format. The provided transcripts have word-level time-stamps on a granularity of 0.1s which allows retrieval systems to index the contents by time offsets. A segment was defined to be a two-minute chunk starting on the minute; e.g. [0.0-119.9] seconds, [60-199.9] seconds, [120-139.9] seconds, etc. Segments overlap each other by one minute - any segment except for the first and last segment is covered by the preceding and following segments. The rationale for creating overlapping segments is to account for the case where a phrase or sentence is split across segment boundaries. This creates 3.4M segments in total from the document collection with an average word count of 340 ± 70 per segment. Topics consist of a topic number, keyword query, a type label, and a description of the user's information need. Eight topics were given at the outset for the participants to practice on, and 50 topics were released as the test task. Topics were formulated in three *types*: topical, re-finding, and known item. Example topics are given in Figure 2.

3.2 Submissions

7 participants submitted 24 experiments for the retrieval task. All runs were 'automatic', i.e, without human intervention; almost all runs were based on the Query Description field, i.e. the more verbose exposition of information need as shown in Figure 2. For training data, many participants used pretrained transfer learning models, some used language technologies and knowledge-based models, and some used only data from the set as shown in table 2. Only one experiment made use of the audio data to produce and use a different transcript than the provided one.

3.3 Evaluation

Two-minute length segments were judged by NIST assessors for their relevance to the topic description. NIST assessors had access to both the ASR transcript (including text before and after the text of the two-minute segment, which can be used as context) as well as the corresponding audio segment. Assessments were made on the PEGFB graded scale (Perfect, Excellent, Good, Fair, Bad) as approximately follows:

Perfect (4): this grade is used only for "known item" and "refinding" topic types. It reflects the segment that is the earliest entry point into the one episode that the user is seeking.

Excellent (3): the segment conveys highly relevant information, is an ideal entry point for a human listener, and is fully on topic. An example would be a segment that begins at or very close to the start of a discussion on the

Participant	run id	field	transfer learning	data processing	IR
Dublin City U	dcu1	D		SpaCy	QE from WordNet
	dcu2	D		SpaCy	QE from Descriptions
	dcu3	D		Spacy	QE, auto RF
	dcu4	D		Spacy	QE from web text
	dcu5	D		Spacy	Combination 1-4
LRG	LRGREtvrs-r_1	D	✓		XLNet;Regression
	LRGREtvrs-r_2	D	✓		XLNet;Regression+Concat
	LRGREtvrs-r_3	D	✓		XLNet;Similarity
U Maryland	UMD_IR_run1	D	✓		Indri
	UMD_IR_run2	D			Indri
	UMD_IR_run3	D	✓	stemming word2vec	Combination + Rerank
	UMD_ID_run4	D	✓	stemming word2vec	rerank + Combination
	UMD_IR_run5	D	✓	stemming	Combination of 1-4
U Texas Dallas	UTDThesis_Run1	D	✓	fuzzy match	Lucene
Johns Hopkins HLT COE	hltcoe1	Q		5-gram	Rocchio RF
	hltcoe2	Q			Rocchio RF
	hltcoe3	Q			no RF
	hltcoe4	D			Rocchio RF
	hltcoe5	Q		transcript 4-gram	Rocchio RF
U Oklahoma	oudalab1	D	✓	SpaCy	BM25; Faiss; finetuned on SQuAD
Spotify	BERT-DESC-S	D	✓		rerank 50; finetuned on other topics
	BERT-DESC-Q	D	✓		rerank 50; finetuned on automatic topics
	BERT-DESC-TD	D	✓		rerank 50; finetuned on synthetic data
baseline	BM25	Q			BM25
	QL	Q			query likelihood
	RERANK-QUERY	Q	✓		rerank 50
	RERANK-DESC	D	✓		rerank 50

Table 2: Technologies employed for the retrieval task

topic, immediately signaling relevance and context to the user.

Good (2): the segment conveys highly-to-somewhat relevant information, is a good entry point for a human listener, and is fully to mostly on topic. An example would be a segment that is a few minutes "off" in terms of position, so that while it is relevant to the user's information need, they might have preferred to start two minutes earlier or later.

Fair (1): the segment conveys somewhat relevant information, but is a sub-par entry point for a human listener and may not be fully on topic. Examples would be segments that switch from non-relevant to relevant (so that the listener is not able to immediately understand the relevance of the segment), segments that start well into a discussion without providing enough context for understanding, etc.

Bad (0): the segment is not relevant.

Figure 3 shows the number of relevant segments of different type per topic. The results are ranged into three groups based on the topic types: topical (15-43), refinding (45-49), known items (53-56). This demonstrates that all topics had some relevant segments retrieved by participants and assessed by assessors.

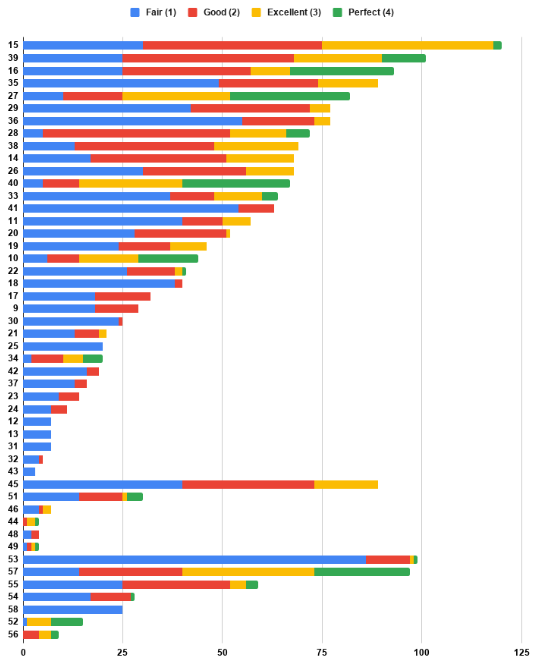


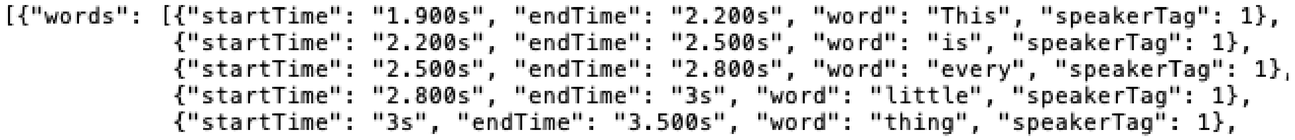
Figure 3: Number of relevant segments of different type per topic, ranged by the number of relevant episodes per three topics categories (topical (15-43), refinding (45-49), known items (53-56).

The primary metric for evaluation is mean nDCG, with normalization based on an ideal ranking of all relevant segments. Note that a single episode may contribute one or more relevant segments, some of which may be overlapping, but these are treated as independent items for the purpose of nDCG computation.

3.4 Search Baselines

Podcast search could be implemented without the full episode transcripts if the titles and creator-provided descriptions provide enough information for search and indexing. As a first baseline, we compared document level retrieval of transcripts to document level retrieval based on titles and creator-provided descriptions. Table 3 shows how using transcripts yields vastly higher scores, compared to using titles or descriptions, episode-level or episode

[2] Implemented using the Pyserini package, `https://github.com/castorini/pyserini` – a Python front end to the Anserini open-source information retrieval toolkit (Yang et al. (2017))



```
[{"words": [{"startTime": "1.900s", "endTime": "2.200s", "word": "This", "speakerTag": 1},
           {"startTime": "2.200s", "endTime": "2.500s", "word": "is", "speakerTag": 1},
           {"startTime": "2.500s", "endTime": "2.800s", "word": "every", "speakerTag": 1},
           {"startTime": "2.800s", "endTime": "3s", "word": "little", "speakerTag": 1},
           {"startTime": "3s", "endTime": "3.500s", "word": "thing", "speakerTag": 1},
```

(a) Transcript snippet

Episode Name	Mini: Eau de Thrift Store
Episode Description	ELY gets to the bottom of a familiar aroma with cleaning expert Jolie Kerr. Guest: Jolie Kerr, of Ask a Clean Person. Thanks to listener Theresa.
Publisher	Gimlet
RSS Link	`https://feeds.megaphone.fm/elt-spot`

(b) Some of the accompanying metadata

Figure 1: Sample from an episode transcript and metadata

	nDCG	nDCG at 30	precision at 10
Episode Title	0.22	0.19	0.12
Episode Description	0.32	0.27	0.17
Episode Title and Description	0.36	0.30	0.19
Episode Title and Description with Show Title and Description	0.37	0.30	0.20
Transcript Text	0.58	0.46	0.41
Transcript Text with Episode Title and Description	0.61	0.49	0.43

Table 3: The contribution of transcripts compared to title search on search results

```
<topic>
<num>34</num>
<query>halloween stories and chat</query>
<type>topical</type>
<description>I love Halloween and I want to hear stories and conversations
             about things people have done to celebrate it.  I am not looking
             for information about the history of Halloween or generalities
             about how it is celebrated, I want specific stories from
             individuals.
</description>
</topic>

<topic>
<num>45</num>
<query>drafting tight ends</query>
<type>refinding</type>
<description>I heard a podcast about strategies for drafting tight ends in
             football.  I'd like to find it again.
</description>
</topic>

<topic>
<num>58</num>
<query>sam bush interview</query>
<type>known item</type>
<description>A bluegrass magazine I read mentioned a podcast interview with
             Sam Bush.  I'd like to hear it.
</description>
</topic>
```

Figure 2: Example search topics

and show-level combined. Even so, adding titles and descriptions to the transcripts improves results somewhat. In all subsequent experiments, the baseline use only transcripts, and do not use show or episode title or descriptions.

Four baseline segment retrieval runs on transcripts are included, using both standard information retrieval methods as well as re-ranking models using BERT (Devlin et al., 2019) to represent segment content.

1. BM25: Standard information retrieval algorithm developed for the Okapi system[2]; the query field of the topic was used for search terms, and up to 1000 segments are returned for each topic.

2. QL (Query Likelihood): Standard information retrieval algorithm[2]

3. RERANK-QUERY: A BERT re-ranking model pre-trained on MS MARCO passage retrieval data (Nogueira and Cho, 2019) without further parameter tuning; the query of the topic was used as the input to the re-ranking model; the re-ranking scores of top-50 segments from BM25 were calculated and submitted per topic.

4. RERANK-DESC: Same as RERANK-QUERY except that the description of the topic was used as the input in re-ranking model.

3.5 Search Results

Table 4 gives an overview of the scores for the submitted experiments. Scoring only the top 30 items or the top 10 items of the list promotes some reranking approaches to the top of the list, illustrating the effect of use case-motivated evaluation metrics on system comparison. One participant resubmitted results after the assessment, to redress the effects of a processing mishap, and those results are marked in the table with an asterisk.

4 Summarization Task

4.1 Definition

Given a podcast episode, its audio, and transcription, the task is to return a short text snippet which accurately conveys the content of the podcast. Returned summaries should be grammatical, standalone utterances of significantly shorter length than the input episode description, short enough to be quickly read on a smartphone screen.

No ground truth summaries are provided; the closest proxies are the show and episode descriptions provided by the podcast creators. We observe that these descriptions vary widely in scope, and are not always intended to act as summaries of the episode content, reflecting the different genres represented in the sample and the different intentions of the creators for the descriptions. We filtered the descriptions to establish a subset that is more appropriate as a ground truth set compared to full set of descriptions. The filtering was done with three heuristics shown in Table 5. These filters overlap to some extent, and remove about a third of the entire set; the remaining 66,245 descriptions we call the *Brass Set*.

4.2 Submissions

8 participants submitted 22 experiments for the summarization task (Table 6). All experiments used some form of deep learning model, and while some used extractive filtering of material from the transcripts as a step in their processes, all were based on abstractive techniques. No participant used the audio data for summary generation.

4.3 Evaluation

The summary labels and scores for the participating systems are created for evaluation sets in two ways.

Manual Assessments and Scoring

Summaries are judged on a four-step scale intended to model how well a listener is able to make a decision whether to listen to a podcast or not, conveying a gist of what the user should expect to

	nDCG	nDCG at 30	precision at 10
UMD_IR_run3	0.67	0.52	0.60
UMD_ID_run4	0.66	0.49	0.56
UMD_IR_run1	0.62	0.45	0.53
UMD_IR_run5	0.65	0.50	0.58
UMD_IR_run2	0.59	0.42	0.51
run_dcu5	0.59	0.43	0.54
run_dcu4	0.58	0.42	0.54
run_dcu1	0.57	0.42	0.50
run_dcu3	0.57	0.42	0.50
run_dcu2	0.55	0.40	0.48
LRGREtvs-r_2 *	0.54	0.40	0.48
LRGREtvs-r_1 *	0.54	0.40	0.47
hltcoe4	0.51	0.43	0.54
LRGREtvs-r_3 *	0.50	0.32	0.41
hltcoe3	0.50	0.35	0.43
hltcoe2	0.47	0.38	0.45
hltcoe1	0.45	0.33	0.38
BERT-DESC-S	0.43	0.47	0.57
BERT-DESC-TD	0.43	0.47	0.56
BERT-DESC-Q	0.41	0.45	0.53
hltcoe5	0.38	0.30	0.37
UTDThesis_Run1	0.34	0.34	0.43
oudalab1	0.00	0.01	0.01
Baseline BM25	0.52	0.40	0.49
Baseline QL	0.52	0.40	0.48
Baseline RERANK-DESC	0.43	0.48	0.57
Baseline RERANK-QUERY	0.43	0.47	0.56

Table 4: Overview of results from submitted segment retrieval experiments. Post-assessment rerun submissions are marked with an asterisk.

filter	criteria	items affected
Length	very long (> 750 characters) or very short (< 20 characters)	24, 033 (23%)
Similarity to other descriptions	> 50% lexical overlap with other episode descriptions	15, 375 (15%)
Similarity to show description	> 40% lexical overlap with own show description	9, 444 (9%)

Table 5: Filters to remove ‘less descriptive’ episode descriptions, to form the *brass subcorpus*.

🟦 = MUKKO 🟪 = BOFLI

Participant	run id	type	method
U New Hampshire 8	unhtrema1	O	GAN, LSTM, 3 sentences, long chunks
	unhtrema2	O	GAN, LSTM, 10 sentences, long chunks
	unhtrema3	O	GAN, LSTM, 20 sentences, short chunks
	unhtrema4	O	GAN, LSTM, 10 sentences, short chunks
U Central Florida 2	UCF_NLP1	A	BART
	UCF_NLP2	A	BART, RoBERTa
U Texas Dallas 7	UTDThesis_Run1	A	T5, fine tuned on brass set + Dialogue Action Tokens
U Glasgow 6	2306987O_abs_run1	A	T5, fine tuned on description
	2306987O_extabs_run2	O	15 sentence input, T5
	2306987O_extabs_run3	O	Extractive filtering, SpanBert
U Cambridge 1	cued_speechUniv1	A	BART, sentence filtering, 9 model ensemble
	cued_speechUniv2	A	BART, sentence filtering, 3 model ensemble
	cued_speechUniv3	A	BART, Fine tuned on transcript
	cued_speechUniv4	A	BART, sentence filtering, non-ensemble
Uppsala U 3	hk_uu_podcast1	A	BART, Longformer, 3 epochs
Spotify 4	categoryaware1	A	BART, Fine tuned on start of transcript + podcast category; 1 epoch
	categoryaware2	A	BART, Fine tuned on start of transcript + podcast category; 2 epochs
	coarse2fine	A	BART, Fine tuned on TextRank center of transcript; 2 epochs
U Delaware 5	udel_wang_zheng1	A	Start of transcript, BART
	udel_wang_zheng2	A	Select sentences by LDA, BART
	udel_wang_zheng3	A	Select sentences by ROUGE, BART
	udel_wang_zheng4	A	Ensemble of 1-3
Baseline	bartcnn	A	BART, No fine tuning
	bartpodcasts	A	BART, Fine tuned on start of transcript
	onemin	E	1 minute of transcript
	textranksegments	E	TextRank, 50 wd segments
	textranksentences	E	TextRank, sentence split

Table 6: Technologies employed for the summarization task

hear listening to the podcast. The assessment scale used by the NIST assessors is the EGFB scale, as per the following instructions:

Excellent: the summary accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. In addition to giving an accurate representation of the content, it contains almost no redundant material which is not needed when deciding whether to listen. It is also coherent, comprehensible, and has no grammatical errors.

Good: the summary conveys most of the most important attributes and gives the reader a reasonable sense of what the episode contains with little redundant material which is not needed when deciding whether to listen. Occasional grammatical or coherence errors are acceptable.

Fair: the summary conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain redundant material which is not needed when deciding whether to listen and may contain repetitions or broken sentences.

Bad: the summary does not convey any of the most important content items of the episode or gives the reader an incorrect or incomprehensible sense of what the episode contains. It may contain a large amount of redundant information that is not needed when deciding whether to listen to the episode.

NIST assessors evaluated 180 of the automatically-generated summaries produced by participants using the EGFB scale. These assessments are converted into a numerical score by a weighting scheme tested for being able separate the baseline systems applied to the Brass set. Weights of 4-2-1-0 for EGFB turned out to be simple and effective in this respect.

In addition to the EGFB assessments, we created a set of boolean attributes that a desirable podcast summary might contain. The primary evaluation metric is the EGFB score; the answers to these attributes are merely an informative signal for participants, and may be useful in devising automated summarization metrics in the future. The attributes are defined from a small-scale survey of podcast listeners, and are listed below.

1. **names:** Does the summary include names of the main people (hosts, guests, characters) involved or mentioned in the podcast?

2. **bio:** Does the summary give any additional information about the people mentioned (such as their job titles, biographies, personal background, etc)?

3. **topics:** Does the summary include the main topic(s) of the podcast?

4. **format:** 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?

5. **title-context:** Does the summary give you more context on the title of the podcast?

6. **redundant:** Does the summary contain redundant information?

7. **english:** Is the summary written in good English?

8. **sentence:** Are the start and end of the summary good sentence and paragraph start and end points?

ROUGE against Creator Descriptions

Each of the test set episodes has a creator-provided description. This description is used as a reference (in the absence of ground truth summaries), and a ROUGE-L Lin (2004) score against the description is computed. ROUGE-L computes overlap of substrings up to the length of the longest common subsequence. Note that these creator-provided descriptions are of varying quality: of the 179-sized subset of descriptions assessed by NIST, we find that only 71, between a third and half, are of Good or Excellent quality.

We provided a version of the episode descriptions processed by a BERT-based sentence classifier that was trained from a small set of manually annotated examples to identify and remove extraneous content such as boilerplate, ads, promotions, and show notes that do not directly summarize or describe the episode (Reddy et al., 2021). 'Cleaned' descriptions with such extraneous content removed were produced, and NIST assessors judged the cleaned descriptions as well as the original descriptions for summary quality.

ROUGE scores were computed against the original episode descriptions, but may be computed against the 'cleaned' descriptions as well.

4.4 Summarization Baselines

Five baseline summarization runs are included. We aimed to include a representation of abstractive as well as extractive models.

1. onemin: Transcript text for the first one minute of the episode.
2. bartcnn: A BART (Lewis et al., 2020) seq2seq model pre-trained on the CNN/Daily Mail corpus for news summarization [3]
3. bartpodcasts: The bartcnn model above, fine-tuned on the full 100k episodes in the dataset, excluding episodes with very short (fewer than 10 characters) or very long descriptions (over 1300 characters). Episodes with descriptions that were highly similar[4] to other descriptions in the same show, or to the show description itself, were also ignored. The descriptions were also processed through a model to detect and remove ads, promotions, and show notes such as links to transcripts.
4. textranksegments: We chunked the transcript into one-minute chunks, and applied the TextRank algorithm (Mihalcea and Tarau, 2004), with word overlap as the similarity metric, to find the most 'central' one-minute segment.
5. textranksentences: The same process as above, except that we chunked the transcript into sentences using SpaCy[5] and extracted the two most central sentences.

4.5 Summarization Results

179 episodes were scored for EGFB quality and the boolean attributes by NIST assessors for the 22 submitted experiments and the 5 baselines. The submitted experiments were also scored automatically using ROUGE-L against the creator-provided descriptions for all the 1024 test episodes. Table 7 shows both the manual assessment scores as well as the automatic evaluations.

All attributes were found to be significantly correlated with the aggregate quality score (Figure 4), to different degrees with 'Does the summary include the main topic(s) of the podcast?' being the most correlated. Future work might investigate these attribute values across all submitted systems towards gaining a concrete understanding of what makes a good podcast summaries.

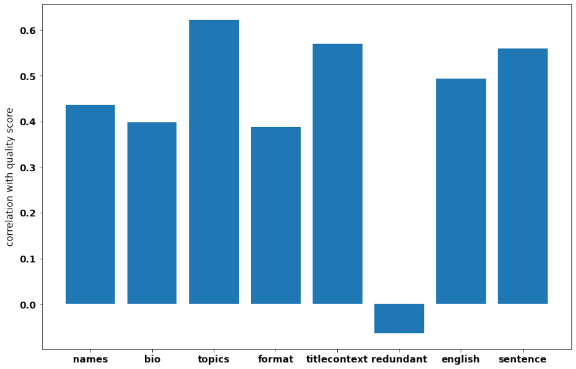


Figure 4: Pearson correlation of attributes with the aggregate EGFB quality score across all submitted baseline runs.

The ROUGE-L F-score is found to be weakly but significantly correlated with the aggregate EGFB

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

[4] We represented each description by a normalized vector of TF-IDF values, and computed similarity as the cosine similarity between the vector representations.

[5] https://spacy.io

quality score (Pearson correlation 0.28). As Figure 5 shows, while summaries rated E and G do have higher median ROUGE scores than those rated F and B, the variation is tremendously large, especially for summaries rated F, raising the question of whether ROUGE against creator descriptions is a sufficiently reliable metric for this task.

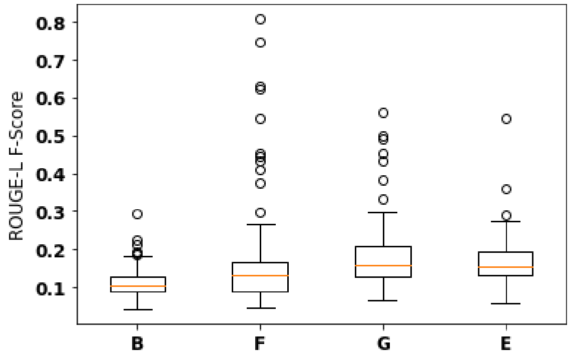


Figure 5: Distribution of ROUGE-L F-Score for each manually assessed label across all submitted baseline runs.

The episodes in the test set were variously challenging. Some very topical podcasts with a clear statement of purpose or a concise topical heading are comparatively easy to summarise, if that statement was identified in the episode transcript or even in the episode description: *"Welcome to my podcast! Let us talk and learn about God's word, life-purpose, values, and faith!"*, or *"On this day in 1826, 15-year-old Ellen Turner was abducted in a forced marriage plot intended to swindle her family out of their fortune."* Episodes with a broad range of covered topics (such as the hosts' opinions on various books, movies and video games and their experiences from working in comedy), and episodes that are not about topics but rather, are sleep aids or avant-garde performance pieces, proved challenging for most systems.

5 Task evolution for Year 2

For 2021, we have the ambition to encourage participants to make use of the audio data in addition to the transcripts, but we do not wish to change the overall task formulation.

We intend to continue the segment retrieval task with some modifications. In 2020, the segment retrieval output was restricted to two-minute segments at fixed starting points over the episode: in 2021, we will consider freely selected jump-in points in the episode, to allow for more precise segment results. We will add new topic types to the three used this year, including types that are likely to be better addressed if the audio signal is taken into consideration.

We will specify the use case which the summarization task is intended to address in greater detail, with the target notion being an Audio Trailer, i.e. the output of the task should be a short highlight clip from the podcast episode in question. In practice, this means that the clip is not required to provide a representation of the entire content but an indicative segment which will inspire the listener to listen to the entire episode. The details of this specification will be formulated to make assessment transparent and reproducible.

Acknowledgments

Building and distributing the dataset and task definitions for the TREC 2020 Podcast Track was a highly collaborative effort, and we have many people to thank for their valuable contributions: Mounia Lalmas-Roelleke, Claire Burke, Zachary Piccolomini, Derek Tang, David Riordan, Fallon Chen, Lex Beattie, Daren Gill, Laura Pezzini, Jenni Lee, Alexandra Wei, Johannes Vuorensola, Nir Zicherman, Max Cutler, Julia Kaplan, Ching-Wei Chen, Brian Brost, Till Hoffmann, Nagarjuna Kumarappan, Maria Dominguez, Laurence Pascall, Md Iftekhar Tanveer, Rezvaneh Rezapour, Hamed Bonab and Jen McFadden.

experiment	aggregate	#E	#E,G	ROUGE-L	ROUGE-L
	EGFB score			recall	precision
cued_speechUniv2	2.04	47	105	0.224	0.235
cued_speechUniv1	1.98	49	104	0.226	0.232
cued_speechUniv4	1.94	46	101	0.204	0.231
UCF_NLP2	1.81	45	92	0.224	0.256
cued_speechUniv3	1.78	39	90	0.205	0.220
hk_uu_podcast1	1.74	35	89	0.190	0.265
UCF_NLP1	1.64	34	79	0.220	0.267
categoryaware2	1.58	32	71	0.199	0.257
categoryaware1	1.51	26	75	0.208	0.227
coarse2fine	1.3	18	57	0.187	0.158
udel_wang_zheng1	1.19	13	52	0.161	0.239
udel_wang_zheng4	1.16	14	53	0.168	0.202
udel_wang_zheng3	1.08	10	44	0.160	0.208
2306987O_abs_run1	1.00	12	39	0.156	0.208
2306987O_extabs_run2	0.99	13	42	0.167	0.237
2306987O_extabs_run3	0.80	8	22	0.147	0.220
udel_wang_zheng2	0.76	7	28	0.139	0.184
UTDThesis1	0.43	1	11	0.129	0.172
unhtrema4	0.04	1	1	0.180	0.069
unhtrema3	0.03	0	0	0.134	0.089
unhtrema2	0.01	0	0	0.090	0.131
unhtrema1	0	0	0	0.061	0.156
Human description	1.45	28	71		
Baseline filtered	1.49	33	71		
Baseline bartpodcasts	1.49	25	75	0.210	0.208
Baseline bartcnn	0.99	10	35	0.272	0.085
Baseline onemin	0.93	5	30	0.282	0.087
Baseline textranksegments	0.38	3	9	0.165	0.083
Baseline textranksentences	0.23	1	4	0.162	0.065

Table 7: Overview of manual assessment results from submitted summarization experiments. The aggregate EGFB score is computed by assigning E=4, G=2, F=1, B=0. ROUGE scores are computed against the original creator provided descriptions of each episode.

References

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. The TREC spoken document retrieval track: A success story (RIAO). In *Content-Based Multimedia Information Access - Volume 1*, pages 1–20, Paris, France, 2000. Le Centre de Hautes Études Internationales d'Informatique Documentaire.

Steve Renals, Thomas Hain, and Hervé Bourlard. Interpretation of multiparty meetings: The AMI and AMIDA projects. In *IEEE Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA 2008)*, 2008.

Alan Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM SIGMM International workshop on Multimedia Information Retrieval*, 2006.

Edison Research. The infinite dial 2020. `https://www.edisonresearch.com/the-infinite-dial-2020/`, 2020. (Accessed on 02/09/2021).

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

Pavel Pecina, Petra Hoffmannová, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. In *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross–Language Evaluation Forum (CLEF 2007). Revised Selected Papers*, 2008.

Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo, and Yoichi Yamashita. Overview of the NTCIR-10 SpokenDoc-2 Task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, 2016.

Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, and Gareth J. F. Jones. Overview of the NTCIR-12 SpokenQuery & Doc-2 Task. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, 2011.

Martha Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth J. F. Jones. Overview of mediaeval 2011 rich speech retrieval task and genre tagging task. In *Working Notes Proceedings of the MediaEval 2011 Multimedia Benchmark Workshop*, 2011.

Maria Eskevich, Gareth J. F. Jones, Shu Chen, Robin Aly, Roeland Ordelman, and Martha Larson. Search and hyperlinking task at mediaeval 2012. In *Working Notes Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*, 2012.

Maria Eskevich, Robin Aly, Roeland Ordelman, David N. Racca, Shu Chen, and Gareth J. F. Jones. SAVA at Mediaeval 2015: Search and anchoring in video archives. In *Working Notes Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*, 2015.

Sebastian Schmiedeke, Peng Xu, Isabelle Ferrané, Maria Eskevich, Christoph Kofler, Martha A. Larson, Yannick Estève, Lori Lamel, Gareth J. F. Jones, and Thomas Sikora. Blip10000: A social video dataset containing spug content for tagging and retrieval. In *Proceedings of the 4th ACM Multimedia Systems Conference*. Association for Computing Machinery, 2013.

Maria Eskevich and Gareth J. F. Jones. Exploring speech retrieval from meetings using the AMI corpus. *Computer Speech & Language*, 28(5), 2014.

Gareth J. F. Jones. About sound and vision: CLEF beyond text retrieval tasks. In *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer, 2019.

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

Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology (JASIST)*, 68(9), 2017.

Jana Besser, Katja Hofmann, and Martha A. Larson. An exploratory study of user goals and strategies in podcast search. In Joachim Baumeister and Martin Atzmüller, editors, *Proceedings from the workshop Lernen, Wissen & Adaptivität (LWA)*. Department of Computer Science, University of Würzburg, Germany, 2008.

Manos Tsagkias, Martha A. Larson, and Maarten de Rijke. Predicting podcast preference: An analysis framework and its application. *Journal of the Association for Information Science and Technology (JASIST)*, 61(2), 2010.

Longqi Yang, Yu Wang, Drew Dunne, Michael Sobolev, Mor Naaman, and Deborah Estrin. More than just words: Modeling non-textual characteristics of podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL)*. Association for Computational Linguistics, 2019.

Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop "Text summarization branches out"*. Association for Computational Linguistics, 2004.

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

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