TREC 2021 Podcasts Track Overview

Jussi Karlgren[1], Rosie Jones[1], Ben Carterette[1], Ann Clifton[1], Maria Eskevich[2], Gareth J. F. Jones[3], Sravana Reddy[4], Edgar Tanaka[1], Md Iftekhar Tanveer[1]

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

Abstract The TREC Podcasts Track is intended to facilitate research in language technologies applied to podcasts specifically by lowering the barrier-to-entry for data-oriented research for podcasts and for diverse spoken documents in general. This year, 2021, is the second year of the track. A more general overview of some of the challenges is given in last year's Podcasts Track Overview. The track this year consisted of two shared tasks: *segment retrieval* and *summarisation*, both based on a dataset of over 100,000 podcast episodes (metadata, audio, and automatic transcripts) which was released concurrently with the track. The tasks were slightly elaborated this year to encourage participants to use audio analysis. This paper gives an overview of the tasks and the results of the participants' experiments.

1 Introduction

The TREC Podcasts Track was launched in 2020 to facilitate research in language technologies applied to podcasts by lowering the barrier-to-entry for data-oriented research for podcasts and for diverse spoken documents in general. A more general overview of some of the starting points is given in the 2021 Podcasts Track Overview (Jones et al., 2021a) and some of the challenges are detailed in separate publications (Jones et al., 2021b; Carterette et al., 2021).

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 in Clifton et al. (2020).

Statistic Name	2020	2021
Email list sign-ups	285	173
Joined TREC slack channel #podcasts-2020	194	67
Registered for TREC podcasts track	213	42
Signed data sharing agreement	77	191
Downloaded transcripts	64	377
Downloaded audio	18	26
Downloaded test audio for summarization	N/A	7
Participated in Search task	7	6
Participated in Summarization task	8	5
Participated in Both tasks	2	1

Table 1: Participation statistics

1.2 Participation

In 2020, the Podcasts Track attracted a great deal of attention with more than 200 registrations to participate. Most registrants did not submit experimental runs for assessment. In 2021, the number of participants who have registered for the Podcast track decreased, while the number of submitted runs stayed comparable, cf. Table 1.

1.3 Tasks

The Podcasts Track offered two tasks: (1) topical retrieval of fixed two-minute segments and (2) textual summarisation of episodes. In 2020, both tasks were possible to complete on the automatic transcripts of episodes, without using the audio data at all. In 2021, we adjusted the tasks somewhat, without changing the overall task formulation, to nudge participants to make more use of the audio material. The segment retrieval, besides the topical retrieval, requested the participants to also rerank the results in additional sets sorted by the segments being entertaining, subjective, or containing discussion of the topic. In addition the two topic types "known-item" and "refinding" from 2020 were merged to "known-item" since we found no practical reason to keep them separate. The summarisation task added the request to submit an audio clip representative of the episode, with slightly more emphasis given in the instructions on the use case of helping a listener decide whether to listen to the episode or not.

2 Segment Retrieval Task

2.1 Previous Work on Retrieval of Spoken Content

There is longstanding interest in spoken content retrieval. TREC organised the Spoken Document Retrieval Track which ran at TREC in the years 1997-2000 (Garofolo et al., 2000) and which focussed on broadcast news. CLEF organised the Cross-Language Speech Retrieval (CL-SR) task which ran at CLEF in the years 2005-2007 (Pecina et al., 2008) and which focussed on retrieval from a large archive of oral history. NTCIR organised a spoken content retrieval task in the years 2010-2016, which focussed on search of Japanese language lectures and technical presentations, including using spoken queries (Akiba et al., 2013; 2016). MediaEval organised the Rich Speech Retrieval and Search and Hyperlinking tasks in the years 2011-2015 which worked with the non-professional video content and the professional broadcast TV material (Larson et al., 2011; Eskevich et al., 2012; 2015).

While none of this existing work has focused on podcast material, 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 issues relating to potential absence of entity mentions in conversational podcasts. 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).

2.2 Task 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. Given a retrieval topic (a phrase, sentence or set of words) and a set of ranking criteria, retrieve and rank relevant two-minute segments from the data.

The provided transcripts have word-level timestamps on a granularity of 0.1s which allows retrieval systems to index the contents by time offsets. A segment is defined to be a two-minute chunk starting on the minute; e.g. [0.0-119.9] seconds, [60-179.9] seconds, [120-239.9] seconds, etc. Segments overlap with 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 the imposed 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 query type, and a description of the user's information need. In 2021, the queries are 40 of type "topical" and 10 of type "known-item". In 2020,

there were 35 of type ”topical”, 8 of type ”refinding”, and 7 of type ”known-item”. Eight topics were given at the outset for the participants to practice on, six of type ”topical”, and one each of ”refinding” and ”known-item”. Example topics are given in Figure 1.

The lists of segments for each topical query are to be submitted in four separately ranked lists: one ranked list of topically relevant segments, and three reranked lists of those same topically relevant segments. Reranking is not relevant for the known-item topics, where the objective is to find one specific segment.[1]

The reranking criteria are:

Adhoc topical retrieval (QR): the segment is topically relevant to the topic description.

Entertaining (QE): the segment is topically relevant to the topic description AND the topic is presented in a way which the speakers intend to be amusing and entertaining to the listener, rather than informative or evaluative.

Subjective (QS): the segment is topically relevant to the topic description AND the speaker or speakers explicitly and clearly express a polar opinion about the query topic, so that the approval or disapproval of the speaker is evident in the segment.

Discussion (QD): the segment is topically relevant to the topic description AND includes more than one speaker participating with non-trivial topical contribution (e.g. mere grunts, expressions of agreement, or discourse management cues (“go on”, “right”, “well, I don’t know …” etc) are not sufficient).

2.3 Submissions

6 participants submitted 23 experiments for the retrieval task. For an overview summary of the submission see Table 2. All runs were ‘automatic’, i.e, without human intervention.

2.4 Evaluation

Submitted two-minute length segments were judged by NIST assessors for their topical relevance to the topic description. Each relevant segment will also be assessed for adherence to the reranking criteria. NIST assessors had access to both the automatically generated 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 intended to be used only for “known item” topics but was used across the board for all topics. It reflects the segment that is the earliest entry point into the intended segment of the intended episode.

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 topic, immediately signalling 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.

[1] We made use of the `Q0` field in the classic TREC retrieval submission format to distinguish between the various reranking schemes. To our knowledge this is the first time that field has been used for anything at all.

```
<topic>
<num>60</num>
<query>is mindfulness effective?</query>
<type>topical</type>
<description>I want information on mindfulness therapy, especially contrasted
with medication or more traditional interventions. Personal testimonials are
relevant as are scientifically oriented ones; mentions of mindfulness in passing
or in a non-therapeutic general language sense are not relevant.</description>
</topic>
...
<topic>
<num>67</num>
<query>pros and cons of ubi</query>
<type>topical</type>
<description>I want to find arguments for and against universal basic
income. </description>
</topic>
...
<topic>
<num>92</num>
<query>edible mushrooms</query>
<type>topical</type>
<description>I would like to know about what mushrooms can be eaten and
enjoyed as food. Names of mushrooms, how to find them, how to prepare
them, how to differentiate between poisonous and edible kinds are all
relevant. Psychoactive mushroom usage and mushrooms in computer
games are not relevant. </description>
</topic>
...
<topic>
<num>103</num>
<query>last emperor of china</query>
<type>known-item</type>
<description>There is a podcast episode featuring a biography of Puyi,
the last emperor of China, that I would like to find.</description>
</topic>
```

Figure 1: Example search topics

Participant	run id	field	transfer learning	IR
TU Vienna	TUW_hybrid_cat	Q	✓	BM25 & $TAS-B-DistilBERT_{DOT}$; re-rank $TAS-B-DistilBERT_{CAT}$
	TUW_hybrid_ws	Q	✓	Hybrid sparse-dense BM25 & $TAS-B-DistilBERT_{DOT}$
	TUW_tasb_cat	Q	✓	$TAS-B-DistilBERT_{DOT}$
	TUW_tasb192_ann	Q	✓	$TAS-B-DistilBERT_{DOT}$ compressed
Open Source Connnections	osc_tok_vec	Q	✓	stemming + SBERT top token hits rescored w/ vector
	osc_vec_tok	Q	✓	SBERT + stemming top vector hits rescored w/ token
	osc_vector	Q	✓	SBERT & pre-trained MS-Marco model
	osc_token	Q		grid tuned BM25F
UCL	UCL_audio_1	D		BM25 + audio feat classifier
	UCL_audio_2	D		BM25 + audio feat rules
Webis	Webis_pc_co_rob	Q	✓	BM25 + Cola & RoBERTa
	Webis_pc_cola	Q	✓	BM25 + Cola
	Webis_pc_rob	Q	✓	BM25 + RoBERTa
	Webis_pc_bs	Q		BM25 baseline
U Waterloo Team h2oloo	tp_mt5	D	✓	BM25 + MS Marco + finetuning
	ms_mt5	D	✓	BM25 + MS Marco
	tp_mt5_f1	D	✓	BM25 + MS Marco + finetuning Yamnet reranking
	tp_mt5_f2	D	✓	BM25 + MS Marco + finetuning Yamnet reranking
U Waterloo Team CFDA	f_coil_tct	D	✓	sparse + dense ensemble: UniCOIL with CLS + TCTColBERT
	f_b25_coil	D	✓	sparse + sparse ensemble: BM25 + UniCOIL with CLS
	f_b25_tct	D	✓	sparse + dense ensemble: BM25 + TCTColBERT
	s_tct	D	✓	dense: TCTColBERT
	s_tasb	D	✓	dense: TAS-B
Baseline	BM25-Q	Q		BM25
	QL-Q	Q		query likelihood
	BM25-D	D		BM25
	QL-D	D		query likelihood

Table 2: Technologies employed for the retrieval task

Bad (0): the segment is not relevant.

The primary metrics for evaluation are mean nDCG, with normalization based on an ideal ranking of all relevant segments, nDCG at the top thirty retrieved items, and precision at ten retrieved items. 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.

2.5 Search Baselines

Four baseline segment retrieval runs on transcripts are included using standard information retrieval methods (BM25 and Query Likelihood, both as implemented in the Pyserini package[2]), each using either the query field only or using both the query and the description fields.

2.6 Relevance Assessment

Figures 2 and 3 show the number of topically relevant segments per test topic for topical queries. This demonstrates that all topics had some relevant segments retrieved by participants and assessed by assessors. Queries are classified by somewhat arbitrarily chosen thresholds as "hard" if less than 20 relevant segments are found among the assessed ones and "easy" if 50 or more are found, as shown in Table 3. The queries for 2021 appear to be somewhat more challenging than the 2020 ones. Table 4 shows the distribution of number of relevant segments over the relevance scores and Table 6 shows the most "hard" and the most "easy" topic of the 2021 training set.

The reranking criteria—*entertaining*, *subjective*, and *discussion*—are variously frequent over the topics. Table 5 shows the number of segments on average per query for the topics, and demonstrates that both "hard" and "easy" topics have on average segments of all three types. Examining the topics individually, we find that whereas several topics have no *entertaining* segments at all, all topics have numerous *subjective* and *discussion* segments. Table 6 gives examples of topics with many or few segments assessed to be *entertaining*, *subjective*, or *discussion*. The examples conform to expectation, in that e.g. indeed one might expect many subjective segments for a topic which explicitly asks for argumentation.

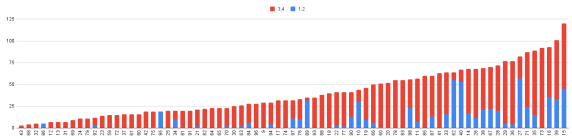

Figure 2: Number of topically relevant segments assessed for both 2020 and 2021 topical queries. Red bars for highly relevant segments (scores 3 and 4); blue bars for less relevant (scores 1 and 2).

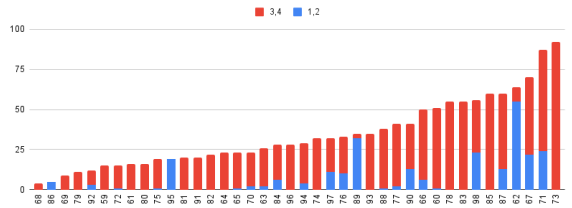

Figure 3: Number of topically relevant segments assessed for 2021 topical queries. Red bars for highly relevant segments (scores 3 and 4); blue bars for less relevant (scores 1 and 2).

2.7 Search Results

Tables 7-10 give an overview of the scores for the submitted experiments for each reranking scheme. 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.

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

	2020	2021	all
Hard (< 20 relevant segments)	9	11	20
Easy (≥ 50 relevant segments)	16	11	26
All	35	40	75

Table 3: Number of hard vs easy queries.

	PEGF	PEG	PE	P
All	1 370	556	257	113
average per query	34	14	6	3

Table 4: Number of Relevant segments among those assessed for 2021 topics.

3 Summarization Task

The user task for summarization is to provide a short description of the podcast episode to help the user decide whether to listen to a podcast. This user task is the background for both the assessment of the text snippet and the audio clip. In particular, participants were required to produce, for a given podcast episode:

1. A short text snippet capturing the most important information in the content of episode, in grammatical utterances of significantly shorter length than the input episode itself.

2. An audio file of up to one minute duration selected from the podcast to give the user a sense of what the podcast sounds like.

3.1 Previous Work on Summarization of Spoken Material

Research in summarization has traditionally focused on text in the news domain (eg Mihalcea and Tarau (2004); Nenkova and McKeown (2011); Hermann et al. (2015)). However, more recently, summarization of other types of content such as dialogues Gliwa et al. (2019) have come into the forefront, and summarization tasks have moved beyond text to encompass transcripts of spoken audio such as meetings Zhong et al. (2021) and media interviews Zhu et al. (2021). By and large, current research relies on human generated transcripts; however, we believe that automatically generated transcripts are a promising input domain. Previous work Spina et al. (2017) has demonstrated that summaries generated using automatically generated transcripts can be comparable in terms of usability to summaries generated using error-free manual transcripts. Modern models like transformers have been shown to be effective at 'correcting'

	Entertaining	Subjective	Discussion
All	457	1081	729
average per query	11	27	18
Hard topics	89	115	99
average per query	8	10	9
Easy topics	118	476	291
average per query	11	43	26

Table 5: Number of Entertaining, Subjective, Discussion segments among those assessed for 2021 topics.

Hard	86	**best drum solo:** I want to hear about great drummers and especially their solos. A segment is relevant if it names the drummer appreciatively and mentions them playing a solo in some song or in some concert.
Easy	62	**descriptions of ponzi schemes and other financial scams:** I want to find stories about Ponzi schemes or similar scams. Segments that claim that something (e.g. bitcoin investing) is a Ponzi scheme or a scam without elaborating on how are not relevant - to be relevant they need to describe the scam and how it works.
Many Entertaining	66	**how to handle failing a job interview:** Any material on how a job interview fails or how a candidate was rejected due to the interview: tips, advice, or personal anecdotes and testimonials are all relevant. If a rejection is not about the interview, even if the segment mentions an interview, it is not relevant.
Few Entertaining	85	**personality disorders:** Discussions about any personality disorder are relevant. Passing claims that someone (e.g. a criminal, a celebrity, the speakers themselves) has a personality disorder without discussing the disorder itself are not.
Many Subjective	67	**pros and cons of ubi:** I want to find arguments for and against universal basic income.
Few Subjective	71	**roman empire:** I am looking to learn something about the history of the Roman Empire
Many Discussion	78	**taboo topics:** I want to understand what topics others consider to be taboo. To be relevant, the segment must mention that a topic is off limits and be clear about what the topic in question is. A mention that some topics are not on is not sufficient. A very general mention that some topics are taboo, e.g. ”sexuality” is partially relevant.
Few Discussion	95	**limericks:** I want to hear limericks. Discussion about limericks is not relevant if a limerick is not included in the segment.

Table 6: Topics with many or few segments assessed to be topically relevant, entertaining, subjective, or discussion

speech recognition errors Hrinchuk et al. (2020), suggesting that summarization models may also have the ability to recover from noisy input.

3.2 Training Data

No ground truth summaries are provided for training or evaluation. The closest proxies are the show and episode descriptions provided by the podcast creators which are included in the released dataset. As described in the 2020 track overview Jones et al. (2021a), these descriptions vary widely in scope, and not all are intended as summaries of the episode. However, most of the submissions using supervised approaches that relied on the creator descriptions as target summaries.

3.3 Submissions

5 participants submitted 11 experiments (Table 11), in comparison to 22 experiments submitted by 8 participants in 2020. All the participants used deep learning, and 3 of the 5 submitted at least one run producing extractive summaries. This is in contrast to 2020, where abstractive summarization dominated – this could be motivated by the additional task to submit a representative audio segment.

As organizers, we provided one baseline: the transcript of first one minute of the episode.

3.4 Evaluation

NIST assessors evaluated 193 of the episodes. Summaries are judged on a four-step scale, as per the following instructions to the assessors.

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.

As in the 2020 task, we devised a set of boolean attributes that a desirable podcast summary might contain.

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 not 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?

Finally, the assessors also gave a binary rating to each submitted audio segment for whether it conveyed a sense of the sound and feel of the podcast episode.

3.5 Summarization Results

Table ?? shows the scores for the 193 assessed episodes. Overall quality scores were significantly lower than in the 2020 task, where the highest mean quality scores were greater than 2.0. Whether that discrepancy is due to the particular test sets, the methods employed, or the annotators is an open question. On the whole, audio segment acceptability scores were high. Consistent with 2020, abstractive systems tended to score higher than extractive ones, though not uniformly so. The first one minute baseline, although simple, proves to be relatively strong.

As in the 2020 task, all attributes were found to be significantly correlated with the aggregate quality score (Figure 4) with 'Does the summary include the main topic(s) of the podcast?' being the most correlated. The audio segment assessment is only weakly correlated with the summary quality.

	egfb	audio	q1	q2	q3	q4	q5	q6	q7	q8
egfb	1.000000	0.224525	0.315758	0.308904	0.621226	0.339426	0.522264	0.293035	0.382040	0.379629
audio	0.224525	1.000000	0.049975	0.085139	0.140858	0.202880	0.110943	0.206044	0.292978	0.262828
q1	0.315758	0.049975	1.000000	0.598698	0.287084	0.301891	0.303172	0.087212	0.122881	0.037317
q2	0.308904	0.085139	0.598698	1.000000	0.231823	0.276614	0.241440	0.086879	0.135384	0.106010
q3	0.621226	0.140858	0.287084	0.231823	1.000000	0.308296	0.660961	0.162771	0.246238	0.230109
q4	0.339426	0.202880	0.301891	0.276614	0.308296	1.000000	0.307444	0.248549	0.287609	0.145186
q5	0.522264	0.110943	0.303172	0.241440	0.660961	0.307444	1.000000	0.140468	0.140593	0.139976
q6	0.293035	0.206044	0.087212	0.086879	0.162771	0.248549	0.140468	1.000000	0.185775	0.111610
q7	0.382040	0.292978	0.122881	0.135384	0.246238	0.287609	0.140593	0.185775	1.000000	0.616005
q8	0.379629	0.262828	0.037317	0.106010	0.230109	0.145186	0.139976	0.111610	0.616005	1.000000

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

Some episodes proved to be easier to summarize than others, with higher aggregate quality scores across systems. Figure 5, the distribution of episode-wise aggregate quality scores, shows that summaries on a minority of episodes get consistently high scores.

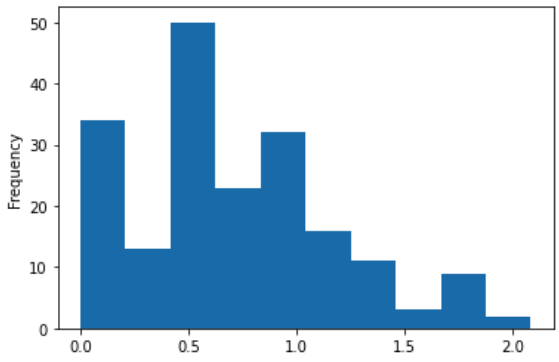

Figure 5: Histogram of EGFB aggregate score averaged across systems per episode. The y axis shows the number of episodes whose average summary quality lies within the given range.

References

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. TREC 2020 Podcasts Track Overview. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC)*. NIST, 2021a.

Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, LongQi Yang, Oguz Semerci, Hugues Bouchard, and Ben Carterette. Current Challenges and Future Directions in Podcast Information Access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021b.

Ben Carterette, Rosie Jones, Gareth F Jones, Maria Eskevich, Sravana Reddy, Ann Clifton, Yongze Yu, Jussi Karlgren, and Ian Soboroff. Podcast metadata and content: Episode relevance and attractiveness in ad hoc search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2247–2251, 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.

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.

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*, 2013.

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*, 2016.

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.

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.

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.

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.

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

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015. URL `http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend`.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*, 2019.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *North American Association for Computational Linguistics (NAACL)*, 2021.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*, 2021.

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.

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7074–7078, 2020. doi: 10.1109/ICASSP40776.2020.9053051.

	nDCG	nDCG at 30	precision at 10
osc_tok_vec	0.39	**0.35**	**0.41**
tp_mt5_f1	**0.51**	**0.37**	**0.40**
tp_mt5_f2	**0.51**	**0.37**	**0.40**
tp_mt5	**0.51**	**0.37**	**0.40**
TUW_hybrid_cat	**0.53**	**0.34**	**0.39**
TUW_hybrid_ws	**0.53**	**0.33**	**0.38**
TUW_tasb_cat	**0.50**	**0.33**	**0.39**
osc_vec_tok	0.35	**0.32**	**0.38**
f_b25_tct	**0.50**	**0.31**	**0.37**
TUW_tasb192_ann	0.43	**0.30**	**0.37**
f_coil_tct	**0.48**	**0.31**	**0.35**
s_tct	0.41	**0.29**	**0.35**
osc_vector	0.31	**0.28**	**0.34**
f_b25_coil	**0.46**	**0.28**	**0.33**
s_tasb	0.39	0.25	**0.32**
osc_token	0.34	**0.28**	**0.32**
ms_mt5	**0.44**	**0.27**	0.28
UCL_audio_2	0.29	0.24	0.28
UCL_audio_1	0.29	0.24	0.28
Webis_pc_bs	0.36	0.20	0.25
Webis_pc_cola	0.36	0.20	0.25
Webis_pc_co_rob	0.36	0.20	0.25
Webis_pc_rob	0.36	0.20	0.25
Baseline BM25-D	0.42	0.25	0.29
Baseline QL-D	0.43	0.25	0.30
Baseline BM25-Q	0.41	0.24	0.27
Baseline QL-Q	0.41	0.25	0.25

Table 7: Overview of results from submitted *topical* (QR) segment retrieval experiments.

	nDCG	nDCG at 30	precision at 10
tp_mt5	**0.30**	**0.20**	**0.18**
TUW_hybrid_cat	**0.31**	**0.16**	**0.15**
osc_vec_tok	0.19	**0.16**	**0.15**
tp_mt5_f1	0.27	**0.16**	**0.14**
osc_tok_vec	0.19	**0.16**	**0.14**
TUW_tasb_cat	0.27	0.15	**0.14**
f_b25_tct	**0.30**	0.15	**0.13**
TUW_hybrid_ws	**0.29**	0.15	**0.13**
f_coil_tct	0.27	0.15	**0.13**
tp_mt5_f2	0.26	**0.16**	**0.12**
ms_mt5	0.27	0.14	**0.12**
s_tct	0.23	0.14	**0.12**
TUW_tasb192_ann	0.23	0.14	**0.12**
osc_vector	0.15	0.12	**0.12**
f_b25_coil	0.27	0.13	0.11
osc_token	0.19	0.15	0.10
Webis_pc_bs	0.23	0.12	0.10
s_tasb	0.22	0.12	0.10
UCL_audio_2	0.15	0.11	0.08
UCL_audio_1	0.15	0.10	0.07
Webis_pc_cola	0.17	0.05	0.05
Webis_pc_rob	0.16	0.04	0.03
Webis_pc_co_rob	0.16	0.03	0.03
Baseline BM25-D	0.24	0.11	0.09
Baseline QL-D	0.25	0.11	0.09
Baseline BM25-Q	0.26	0.14	0.10
Baseline QL-Q	0.27	0.14	0.10

Table 8: Overview of results from submitted *entertaining* (QE) segment retrieval experiments.

	nDCG	nDCG at 30	precision at 10
tp_mt5	**0.47**	**0.31**	**0.35**
tp_mt5_f1	**0.46**	**0.31**	**0.33**
f_b25_tct	**0.49**	**0.29**	**0.32**
osc_vec_tok	0.32	**0.28**	**0.32**
osc_tok_vec	0.31	**0.27**	**0.32**
TUW_tasb192_ann	0.39	**0.25**	**0.32**
tp_mt5_f2	**0.46**	**0.30**	**0.31**
f_coil_tct	**0.46**	**0.28**	**0.31**
s_tct	0.41	**0.27**	**0.29**
TUW_hybrid_ws	**0.48**	**0.27**	**0.29**
TUW_tasb_cat	**0.46**	**0.27**	**0.28**
f_b25_coil	**0.45**	**0.24**	**0.28**
TUW_hybrid_cat	**0.49**	**0.26**	**0.27**
osc_vector	0.28	0.23	**0.27**
s_tasb	0.37	0.22	**0.27**
ms_mt5	0.41	0.22	0.23
osc_token	0.27	0.21	0.23
UCL_audio_1	0.25	0.19	0.20
UCL_audio_2	0.24	0.19	0.20
Webis_pc_bs	0.34	0.17	0.20
Webis_pc_cola	0.24	0.06	0.06
Webis_pc_co_rob	0.23	0.04	0.06
Webis_pc_rob	0.22	0.04	0.04
Baseline BM25-D	0.40	0.21	0.24
Baseline QL-D	0.41	0.21	0.24
Baseline BM25-Q	0.40	0.22	0.22
Baseline QL-Q	0.41	0.22	0.21

Table 9: Overview of results from submitted *subjective* (QS) segment retrieval experiments.

	nDCG	nDCG at 30	precision at 10
f_b25_tct	**0.43**	**0.24**	**0.22**
f_coil_tct	**0.41**	**0.24**	**0.22**
osc_tok_vec	0.27	**0.22**	**0.22**
tp_mt5	**0.39**	**0.22**	**0.21**
tp_mt5_f1	0.38	**0.22**	**0.21**
tp_mt5_f2	0.38	0.21	**0.21**
TUW_tasb192_ann	0.34	0.20	**0.21**
TUW_hybrid_ws	**0.44**	**0.24**	0.20
osc_vec_tok	0.28	**0.23**	0.20
f_b25_coil	**0.40**	0.21	0.20
s_tct	0.35	**0.22**	0.19
TUW_tasb_cat	**0.39**	0.21	0.19
s_tasb	0.34	0.20	0.19
osc_vector	0.24	0.19	0.19
TUW_hybrid_cat	**0.42**	0.21	0.18
ms_mt5	0.36	0.18	0.17
osc_token	0.25	0.19	0.16
Webis_pc_bs	0.32	0.16	0.16
UCL_audio_2	0.25	0.19	0.15
UCL_audio_1	0.24	0.19	0.15
Webis_pc_cola	0.23	0.06	0.06
Webis_pc_co_rob	0.22	0.05	0.06
Webis_pc_rob	0.21	0.04	0.04
Baseline BM25-D	0.36	0.19	0.18
Baseline QL-D	0.37	0.18	0.19
Baseline BM25-Q	0.37	0.20	0.16
Baseline QL-Q	0.37	0.20	0.15

Table 10: Overview of results from submitted *discussion* (QD) segment retrieval experiments.

participant	run IDs	type	method
PoliTO	PoliTO_100_32-128, PoliTO_25_32-128, PoliTO_50_32-128, PoliTO_50_64-128	Abstractive	Extractive filtering: A sentence-BERT model to obtain the embeddings for each sentence of the podcast, and a fully connected supervised model to fuse text and audio features (obtained using OpenSmile by the track organizers). Abstractive summary generation: A LongFormer (LED) model fine-tuned on creator descriptions.
Webis	Webis_pc_abstr	Abstractive	Roberta Model for transfer learning, own annotations, DistilBART
	Webis_pc_extr	Extractive	Roberta Model for transfer learning, own annotations, SentenceBERT for sentence similarity
theTuringTest	theTuringTest1	Extractive	Feature engineering, metrics such as Rouge1, Rouge2, RougeL, and Meteor against creator descriptions, TOPSIS
	theTuringTest2	Abstractive	Feature engineering, metrics such as Rouge1, Rouge2, RougeL, and Meteor against creator descriptions, T5
UniCamp	Unicamp1, Unicamp2	Abstractive	mBART adapted to use Longformer attention, trained on podcast transcripts in English and Portuguese
Spotify	Hotspot1	Extractive	Speech emotion recognition model trained with external resources to produce the representative audio segments. SentenceBERT + centrality to select the extractive summary.
Baseline	onemin	Extractive	1 minute of transcript

Table 11: Technologies employed for the summarization task

experiment	type	quality	audio
PoliTO_50_64-128	A	1.06	0.98
Unicamp1	A	1.04	1.00
PoliTO_25_32-128	A	1.03	0.98
Unicamp2	A	1.01	0.50
PoliTO_100_32-128	A	0.98	0.99
PoliTO_50_32-128	A	0.91	0.99
Baseline onemin	E	0.81	0.96
Hotspot1	E	0.43	0.95
theTuringTest1	E	0.34	0.20
Webis_pc_extr	E	0.26	0.92
Webis_pc_abstr	A	0.23	0.94
theTuringTest2	A	0.18	0.21

Table 12: Overview of manual assessment results from submitted summarization experiments. A denotes abstractive and E extractive systems. The quality score is aggregated from the EGFB assessments by assigning E=4, G=2, F=1, B=0 and averaging; i.e., the scale of the quality score is from 0 to 4. The audio segment acceptability is a binary assessment where the scale is from 0 to 1.