100,000 Podcasts: A Spoken English Document Corpus

Fabio Viggiano

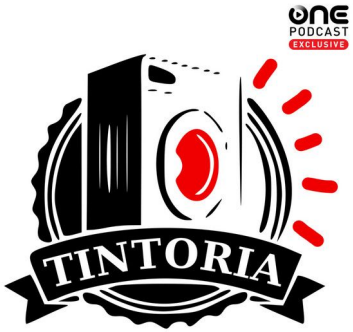
fabio.viggiano@studio.unibo.it

LM Informatica

2024/2025

Introduction

Podcast: A daily routine for many people.

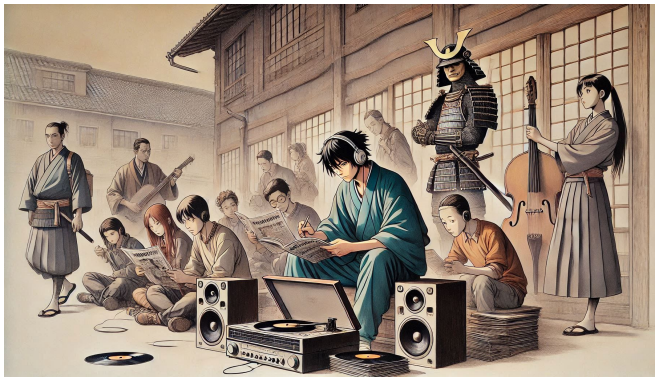


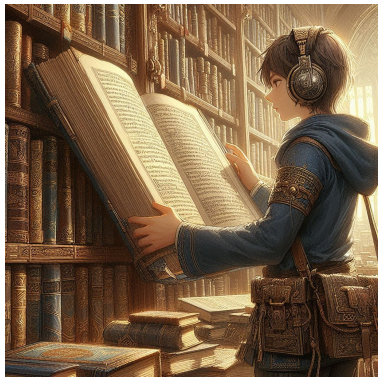
The challenges

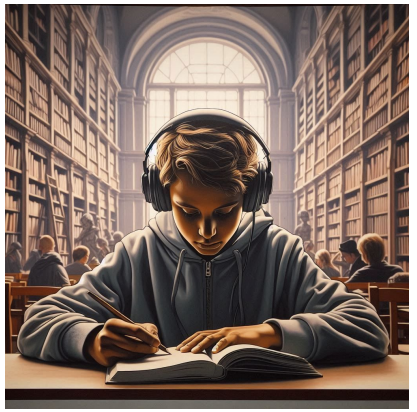
Automatic speech recognition

Passage search

Summarization







Natural language processing, information retrieval and linguistics

Related datasets

Clean audio

TIMIT collection (Garofolo et al., 1990)
TREC (Garofolo et al., 2000)
CLEF (Federico e Jones, 2003)

These more formal settings or samples of formal content are useful for the study of acoustic qualities of human speech, but represent a more idealized scenario than practical audio processing tasks of interest today.

Conversational datasets with noisier speech

ATIS corpus of air travel - Information requests (Hemphill et al., 1990),
Meeting recordings (Garofolo et al., 2004b),
Conversations (Canavan et al., 1997; Godfrey and Holliman, 1993),
Broadcast news (Garofolo et al., 2004a).

Conversational datasets with noisier speech have been collected for specific domains, often intended to capture regularities of some particular communication situation.

Summarization datasets

AMI meeting corpus (Mccowan et al., 2005)
ICSI meeting corpus (Janin et al., 2003)
Corpora of lectures (Miller, 2019)

Summarization datasets with manually written summaries.

Spotify Podcast Dataset

Spotify has created a corpus on the topic that includes almost **60,000 hours** of speech.

- the **largest corpus of transcribed speech data**,
- a set of **labeled data** on this corpus,
- **benchmarking results** using standard baselines,
- an **analysis of the data and benchmarking results**
- advanced speech research, which **fills the gaps in existing datasets**

The characteristics of the dataset

Language: English

Length: filtered out any non-professionally published episodes that are longer than 90 minutes

Speech Presence: ignored episodes that are less than 50% speech over the duration of the episode.

Episodes: over 100’000

18,376 podcast shows

Average episode duration is **33.8 minutes**

Creator-provided episode descriptions average **85 words** in length

Common categories: **comedy**, **sports**, **health & fitness**, **society & culture**, **science**, **news** and **politics**

Geographic origins: **US**, **Great Britain**, **Canada**, **Australia** and **India**

Median speaker turn length per episode is about **110 seconds**

Text transcripts

The text transcripts have been generated by Google's Cloud Speech-to-Text API 2 with:

- word-level time alignments for each word
- speaker diarization
- punctuation

An example snippet

Transcript and metadata


```
[{"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

Sample word error rate of 18.1% and a Named entity recognition accuracy of 81.8%.

Search: Spoken Passage Retrieval

Challenge: Searching for specific content within podcasts.

Task: The case study in this article focuses on the retrieval of fixed-length segments. Given an input (a phrase, sentence, or set of words), the relevant text segments are retrieved.

Search: Spoken Passage Retrieval

The case-study is for **fixed-length segment** retrieval: given an arbitrary query (a phrase, sentence or set of words), retrieve topically relevant segments from the data.

A segment, for the purposes of our benchmark, is a **two-minute chunk** with one minute overlap and starting on the minute.

This creates **3.4M segments** in total from the benchmark with the average word count of 340 ± 70.

Evaluation Data for Search

Research information called "**topics**" was created based on those used by the Text REtrieval Conference (**TREC**).

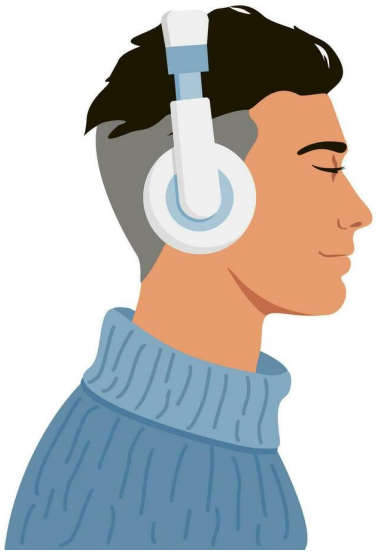
Each topic consists of a **keyword query** and a **description of the user's information need**.

Examples of topics:

- topical (general information)
- re-finding (find an episode already listened to)
- known item (finding something known but with an unknown name)

Human judgment

Gold standard data for evaluation consists of human judgments of the relevance of segments to the topics.



Excellent
Good
Fair
Bad

Basis of comparison

- Baseline system

BM25-based search (Okapi BM25): a ranking function used by search engines to estimate **the relevance of documents to a given search query**. The similarity score is always between 0 and 1 and the parameters are:

*K1 - controls the **saturation of the frequency** of the term*
*B - controls how much effect **document length normalization** should have*

It has been used to retrieve segments for judging, manually varying the query terms to try to increase coverage.

Query Likelihood (QL**):** A probabilistic model that estimates the probability of a document given a query.

- Relevance feedback

RM3: A relevance model used for relevance feedback, improving search results.

Implementation details

- **Pyserini** - An open-source search toolkit used for implementing the search functionality.
- **Lucene** - The underlying search library used by Pyserini.
- **Porter** - A specific stemming algorithm, known for its simplicity and effectiveness.

Models evaluated

- BM25
- BM25 with RM3 relevance feedback (BM25+RM3)
- QL (Query Likelihood)
- QL with RM3 relevance feedback (QL+RM3)

nDCG evaluation

Discounted cumulative gain (DCG) is a measure of ranking quality in information retrieval.

$$\text{NDCG@K} = \frac{DCG@K}{IDCG@K}$$

NDCG = 1	in the case of ideal ranking when items are perfectly sorted by relevance.
NDCG = 0	when there are no relevant objects in top-K.
0 < NDCG < 1	in all other cases

The higher the NDCG, the better

Results for search

		nDCG@5				nDCG@10			
		BM25	BM25+RM3	QL	QL+RM3	BM25	BM25+RM3	QL	QL+RM3
1	coronavirus spread	0.6655	0.6597	**0.7169**	0.5933	0.6717	**0.7278**	0.678	0.6579
2	greta thunberg cross atlantic	0.5801	0.1461	**0.8136**	0.4469	0.4742	0.2731	**0.5655**	0.391
3	black hole image	**0.8721**	0.851	0.7261	0.7104	**0.7921**	0.785	0.7325	0.7413
4	story about riding a bird	0	0	0	0	0	0	0	0
5	daniel ek interview	0	0	0	0	0	0	0	0
6	michelle obama becoming	**0.0838**	0	0	0	**0.0643**	0	0.0363	0
7	anna delvey	0	0	0	0	0	0	0	0
8	facebook stock prediction	0.5591	0.3367	**0.7016**	0.4409	0.6005	0.5394	**0.6792**	0.5477
	all	0.3451	0.2492	**0.3698**	0.2739	0.3253	0.2907	**0.3364**	0.2922

Table 5: nDCG scores for 8 human expert annotated topics.

	nDCG@5	nDCG@10
BM25	0.2737	0.3325
BM25+RM3	0.2731	0.3261
QL	0.2660	0.3357
QL+RM3	0.2542	0.3329

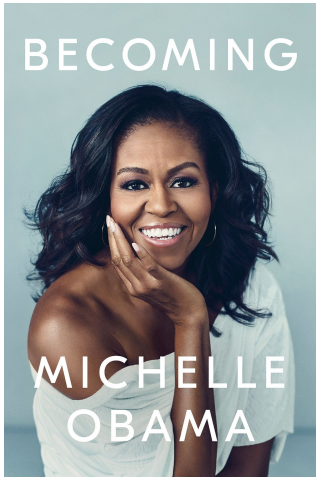
Table 6: nDCG scores for 14 crowdsourced test topics.

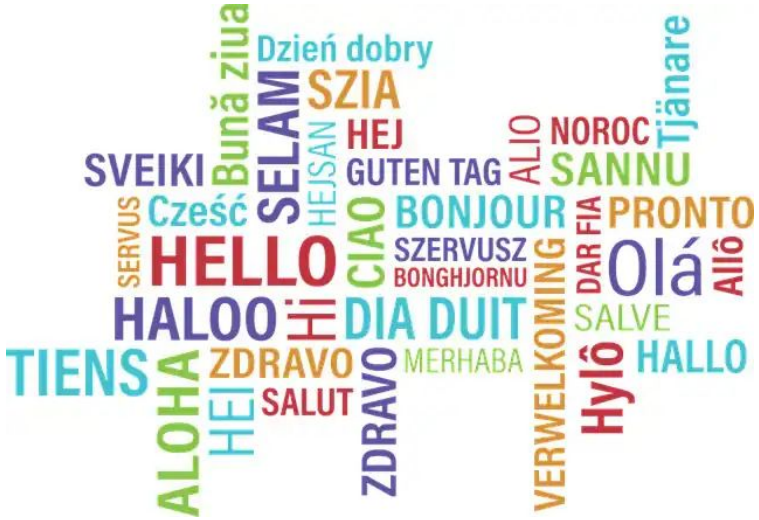
QL better than BM25

Lessons Learned for Spoken Passage Retrieval

The evaluation of IR systems on podcast transcripts revealed several limitations:

- **Basic Bag-of-Words Approach**
- **Automatic Speech Recognition Errors**
- **Query-Term Matching Limitations**
- **Language Diversity**





Summarization

Automated document summarization is the task of condensing an input text into a much shorter form that preserves most of the salient information

This dataset presents several challenges:

- **speech recognition errors,**
- **conversational nature,**
- **the documents are significantly longer than typical summarization data.**

Data Preparation for Summarization

Brass Subcorpus

The creator-generated descriptions have been considered as reference summaries, but these have been filtered in order to select a subset of the corpus that is suitable for training supervised models.

Length	descriptions that are very long (> 750 characters) or short (< 20 characters) amounting to $24,033$ or 23% of the descriptions.
Similarity to other descriptions	descriptions with high lexical overlap (over 50%) with other episode descriptions amounting $15,375$ or 15% of the descriptions.
Similarity to show description	descriptions with high lexical overlap (over 40%) with their show description, amounting to $9,444$ or 9% of the descriptions.

Table 7: Filters to remove less descriptive episode descriptions, to form the *brass subcorpus*.

Data Preparation for Summarization

Filters

The descriptions were filtered using three heuristics:

- length,
- similarity to other descriptions
- similarity to the descriptions of the programs they posted to

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**.

Data Preparation for Summarization

Gold Test Data

To derive gold labeled data, the outputs have been labeled of different baseline systems on a sample of 303 episodes.

The researchers asked annotators to assess a summary's quality on a EGFB scale, after reading the full transcript and/or listening to some of the audio if needed.

Baseline Systems

Unsupervised Extractive Summary

The researchers employed **TextRank**, an unsupervised summarization technique, to identify the most important sentences in the test data.

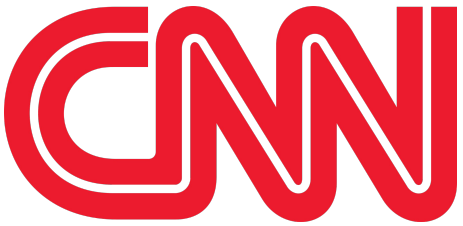
This method constructs a graph of sentences, with edges representing their similarity.

PageRank is then used to determine the most central sentences, which are treated as the summary.
Additionally, a simple baseline was established by using **the first minute** of spoken content for comparison.

Baseline Systems

Supervised Extractive

We ran two variants of supervised models for generating **abstractive summaries**, both using **BART (** a denoising autoencoder for pretraining sequence-to-sequence models)



Evaluating Summary Quality

Comparison between Automatic Metrics (ROUGE) and Human Evaluations of Summaries

	Brass			Non-Brass		
	R1-F	R2-F	RL-F	R1-F	R2-F	RL-F
FIRST MINUTE	18.90	3.92	9.68	16.89	3.67	9.78
TEXTRANK	15.25	2.04	8.69	13.04	1.58	7.99
BART-CNN	20.67	4.87	12.6	22.93	5.3	14.52
BART-PODCASTS	28.24	13.34	21.39	29.46	12.87	22.07

Table 8: ROUGE scores bucketed by whether the test descriptions passed the brass filter.

ROUGE used to automatically evaluate the lexical overlap of the summaries generated by the models compared to the descriptions provided by the creators.

	Brass					Non-Brass				
	E	G	F	B	Pct Good or Better	E	G	F	B	Pct Good or Better
CREATOR	37	35	33	39	50%	36	57	36	30	58%
FIRST MINUTE	10	33	46	55	30%	10	38	61	50	30%
TEXTRANK	2	10	43	89	8%	3	13	35	108	10%
BART-CNN	8	47	40	49	38%	28	40	41	50	43%
BART-PODCASTS	17	50	43	34	47%	37	51	35	36	55%

Table 9: Human labeled score distribution

Human analysts assessed the quality of the summaries by comparing them not with descriptions (which are often of low quality), but directly with podcast transcripts

The supervised systems BART-CNN and BART-PODCASTS have comparable results to creator summaries. Unsupervised systems return the lowest performance.

Rogue details

BART-PODCASTS emerges as the best performing model, from a point of view of lexical overlap with the creators' descriptions, the summaries produced by this model are the most similar.

The ROUGE results are, overall, lower than summary benchmarks on news texts.

Analysis of Summarization Results

- **Creator descriptions are not great summaries**

  While used as a baseline, creator descriptions are inconsistent in quality.

- **BART-PODCASTS model performs well**

  This model, specifically fine-tuned for podcasts, performs well on both a "brass labeled" set (expert-labeled) and a non-brass set.

Analysis of Summarization Results

- **ROUGE scores are reliable in this case**

  Although ROUGE scores (automatic evaluation metric) can be unreliable for spoken language, **they seem to correlate well with human judgment in this case**.

  This is further supported by the model ranking consistency across different reference summary qualities (good vs. bad).

- **Abstractive models outperform extractive ones**

  Models that can generate new text (abstractive) perform better than models that just pick existing sentences (extractive). This is because abstractive models can overcome speech recognition errors and generate fluent summaries.

Conclusions and Future Work

In the NLP domain, podcasts are an ideal test not only for **retrieval** and **summarization from transcripts**, but also end to end **summarization-translating** the original audio into either a written summary or a short audio trailer – or **retrieval tasks** that leverage the audio, such as **keyword search** and s**poken document retrieval**.

The very varied styles and topics in the corpus suggest that this data may be of interest to research in **sociolinguistics** or **computational social science**.

References

[1] “100,000 Podcasts: A Spoken English Document Corpus” by Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones, COLING 2020

https://www.aclweb.org/anthology/2020.coling-main.519/

[2]”BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension” by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer 2020

https://arxiv.org/pdf/1910.13461