The English Language Podcast Dataset, Research Questions, and the TREC Podcasts Challenge

One New Dataset, Two New Shared Tasks, and Many New Interesting Questions

Jussi Karlgren | 18 feb 2021 | Helsinki

Spotify

jussi karlgren

- 1990-2010 SICS
  - computational stylistics, interaction, information retrieval evaluation
- 2010-2019 Gavagai
  - semantic spaces, sentiment analysis, media monitoring
- 2019- Spotify
  - podcasts

and every now and then, in between, some other interesting places:

- 1997-1999 Helsinki
  - acting professor, computational linguistics

very much a team effort

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich (CLARIN EU), Gareth J.F. Jones (Dublin City U), Jussi Karlgren, Aasish Pappu, Sravana Reddy, Yongze Yu

Spotify

we have released a data set of podcasts for research purposes

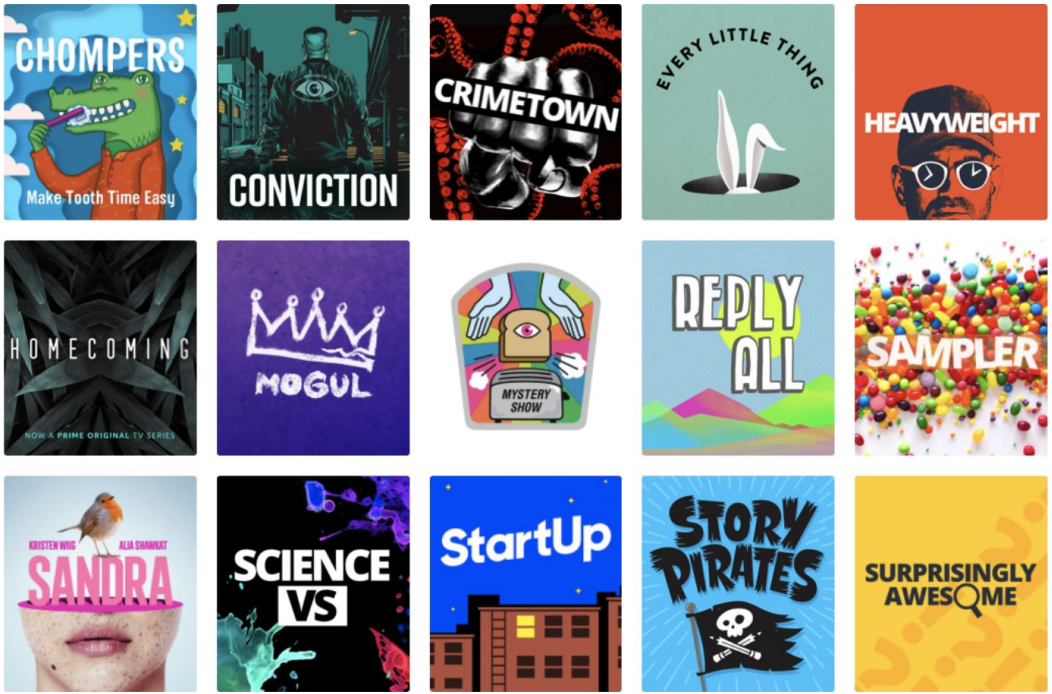
through TREC, with two challenge tasks:

- segment retrieval
- summarisation

https://podcastsdataset.byspotify.com/

Spotify

A New English-Language Corpus!



The first fully-transcribed, large-scale podcast dataset

2TB of data - 100k episodes with audio

Spotify

This podcast corpus is vastly larger than previous speech datasets.

Switchboard: ~110 h

TDT-2 Corpus (TREC-9 SDR) ~600 h

Fisher corpus: ~2k h

Spotify Podcast Corpus: ~50k h

Spotify

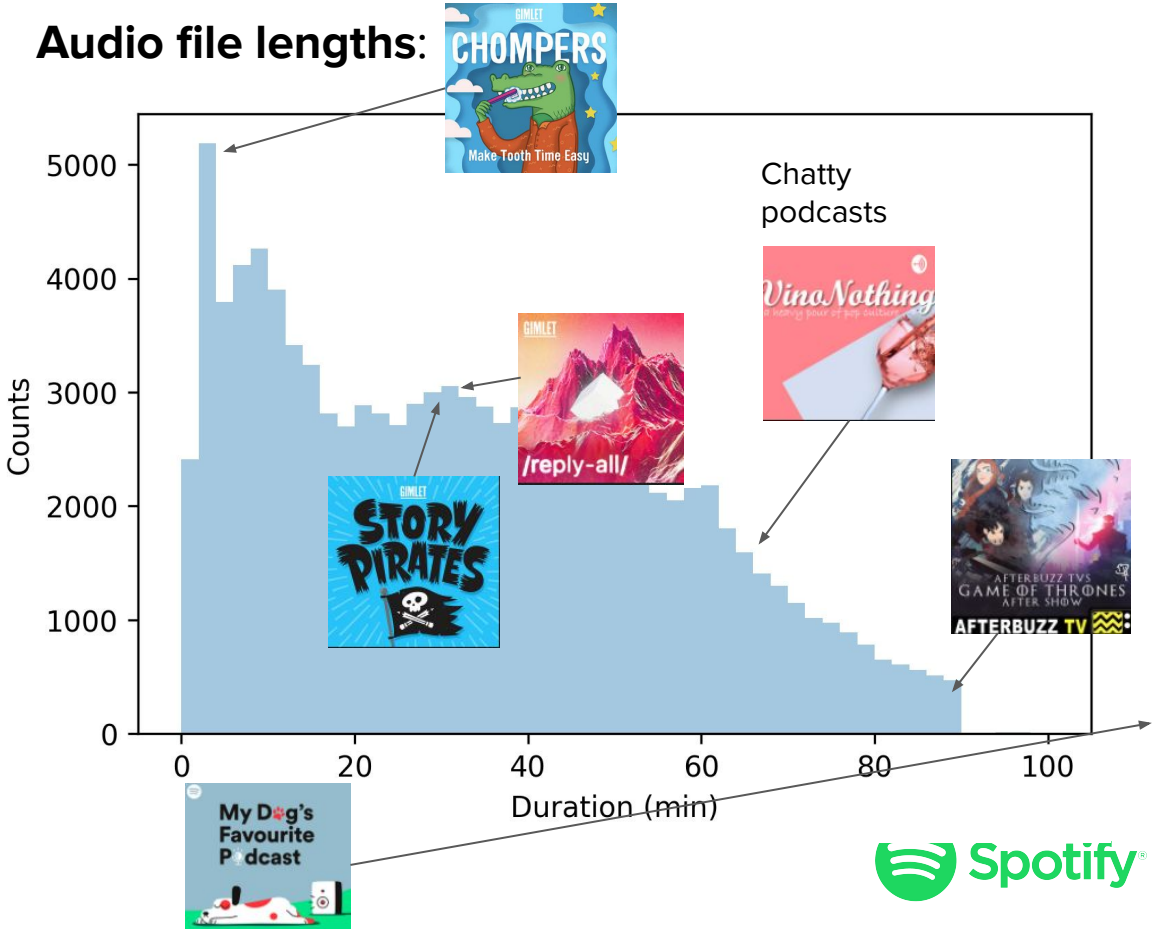
Dataset By the Numbers

File Sizes

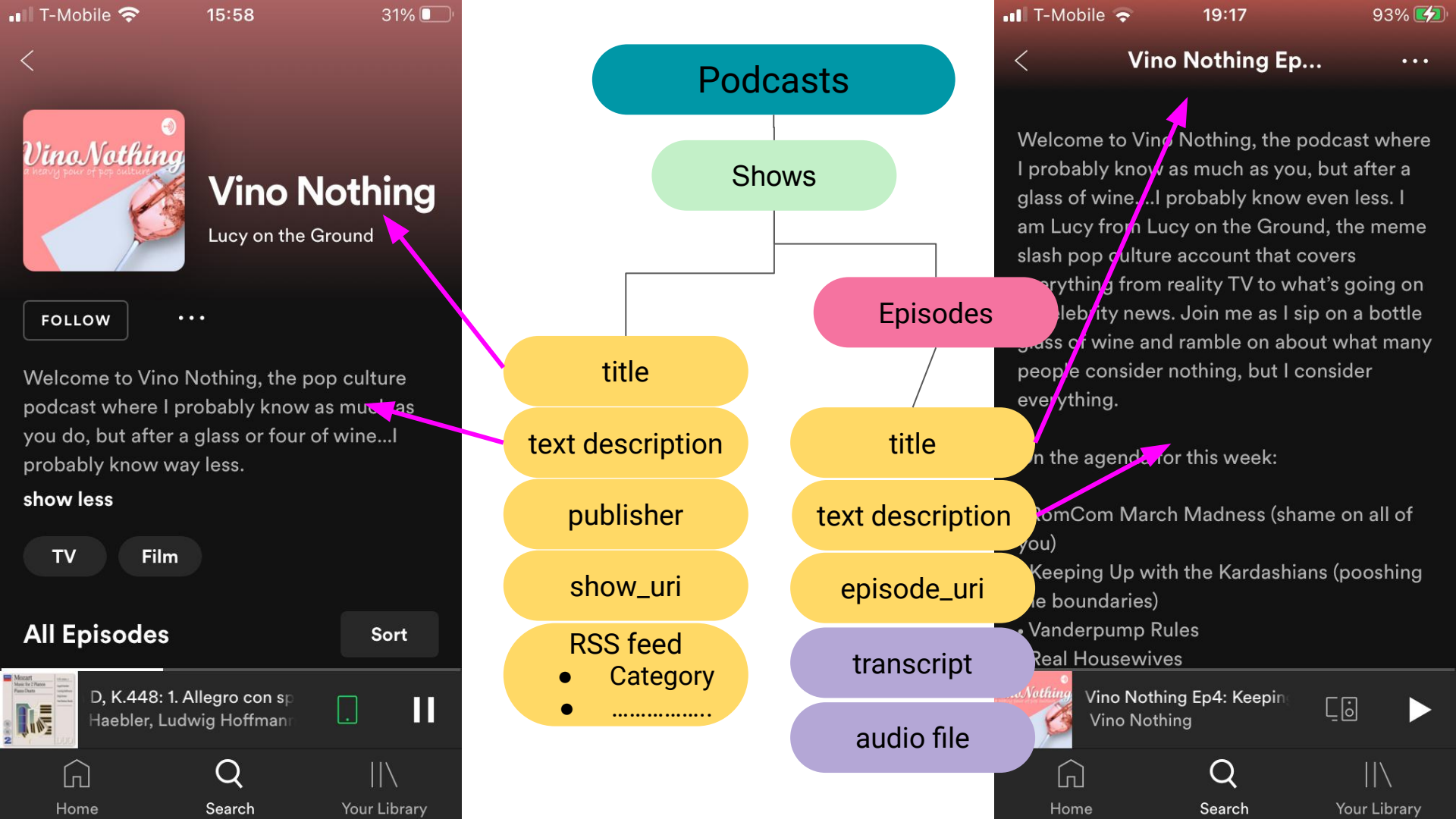
- ~ 100 000 episodes
- Transcripts and metadata: 13 GB
- Audio files: 2 TB

Transcript word counts:

- avg: 5 728
- max: 43 504



the data set is structured


T-Mobile
15:58
31%
Vino Nothing
Lucy on the Ground
FOLLOW
Welcome to Vino Nothing, the pop culture podcast where I probably know as mu as you do, but after a glass or four of wine...I probably know way less.
show less
TV
Film
All Episodes
Sort
D, K.448: 1. Allegro con sp
Haebler, Ludwig Hoffmanr
Home
Search
Your Library
Podcasts
Shows
Episodes
title
text description
publisher
show_uri
RSS feed
Category
.....
title
text description
episode_uri
transcript
audio file
T-Mobile
19:17
93%
Vino Nothing Ep...
Welcome to Vino Nothing, the podcast where I probably know as much as you, but after a glass of wine...I probably know even less. I am Lucy from Lucy on the Ground, the meme slash pop culture account that covers everything from reality TV to what's going on celebrity news. Join me as I sip on a bottle glass of wine and ramble on about what many people consider nothing, but I consider everything.
In the agenda for this week:
RomCom March Madness (shame on all of you)
Keeping Up with the Kardashians (pooshing the boundaries)
Vanderpump Rules
Real Housewives
Vino Nothing Ep4: Keepin
Vino Nothing
Home
Search
Your Library

Transcripts were generated using **third-party high-quality ASR**


```
[{"words": [{"startTime": "0.900s", "endTime": "1.400s", "word": "Welcome", "speakerTag":
1}, {"startTime": "1.400s", "endTime": "1.500s", "word": "to", "speakerTag": 1},
{"startTime": "1.500s", "endTime": "1.700s", "word": "the", "speakerTag": 1},
{"startTime": "1.700s", "endTime": "2.100s", "word": "AIA", "speakerTag": 1},
{"startTime": "2.100s", "endTime": "2.800s", "word": "Vitality", "speakerTag": 1},
{"startTime": "2.800s", "endTime": "3.100s", "word": "One", "speakerTag": 1},
{"startTime": "3.100s", "endTime": "3.400s", "word": "Minute", "speakerTag": 1},
{"startTime": "3.400s", "endTime": "4.200s", "word": "Podcast.", "speakerTag": 1},
```


Spotify

Transcripts were generated using **third-party high-quality ASR**

Features include:

- Diarization (speaker tagging)


```
[{"words": [{"startTime": "0.900s", "endTime": "1.400s", "word": "Welcome", "speakerTag":
1}, {"startTime": "1.400s", "endTime": "1.500s", "word": "to", "speakerTag": 1},
{"startTime": "1.500s", "endTime": "1.700s", "word": "the", "speakerTag": 1},
{"startTime": "1.700s", "endTime": "2.100s", "word": "AIA", "speakerTag": 1},
{"startTime": "2.100s", "endTime": "2.800s", "word": "Vitality", "speakerTag": 1},
{"startTime": "2.800s", "endTime": "3.100s", "word": "One", "speakerTag": 1},
{"startTime": "3.100s", "endTime": "3.400s", "word": "Minute", "speakerTag": 1},
{"startTime": "3.400s", "endTime": "4.200s", "word": "Podcast.", "speakerTag": 1},
```


Spotify

Transcripts were generated using **third-party high-quality ASR**

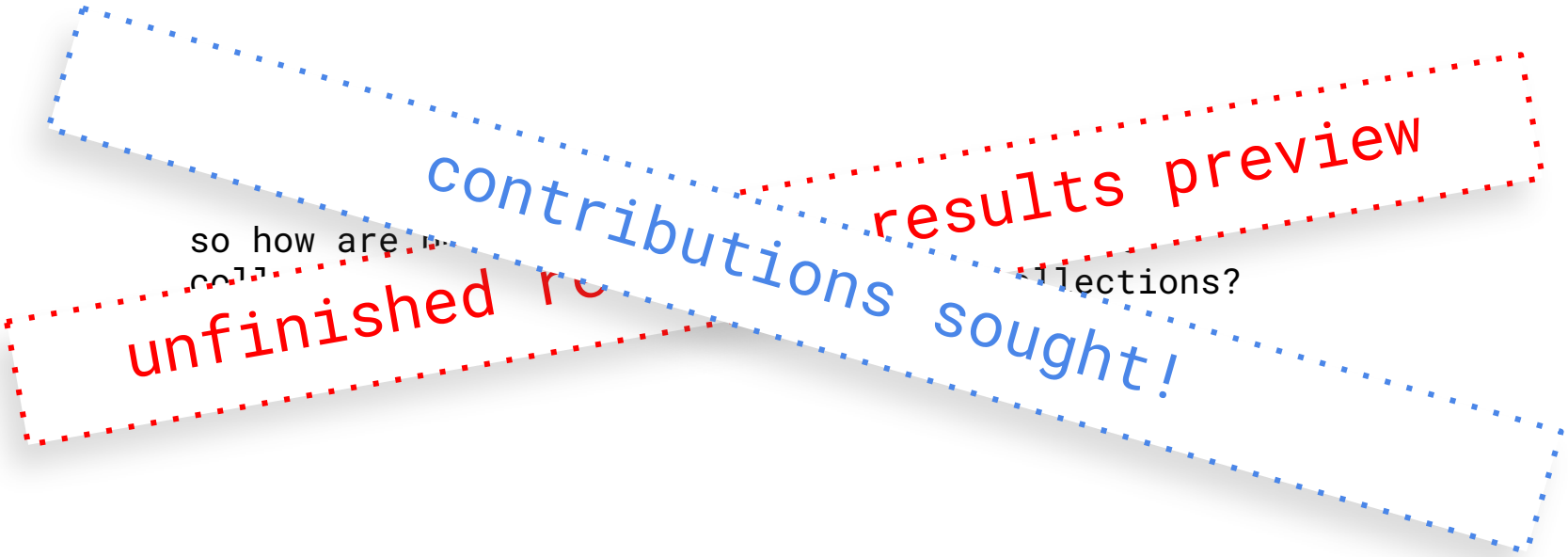
Features include:

- Diarization (speaker tagging)
- Inferring punctuation/sentence segmentation


```
[{"words": [{"startTime": "0.900s", "endTime": "1.400s", "word": "Welcome", "speakerTag":
1}, {"startTime": "1.400s", "endTime": "1.500s", "word": "to", "speakerTag": 1},
{"startTime": "1.500s", "endTime": "1.700s", "word": "the", "speakerTag": 1},
{"startTime": "1.700s", "endTime": "2.100s", "word": "AIA", "speakerTag": 1},
{"startTime": "2.100s", "endTime": "2.800s", "word": "Vitality", "speakerTag": 1},
{"startTime": "2.800s", "endTime": "3.100s", "word": "One", "speakerTag": 1},
{"startTime": "3.100s", "endTime": "3.400s", "word": "Minute", "speakerTag": 1},
{"startTime": "3.400s", "endTime": "4.200s", "word": "Podcast", "speakerTag": 1},
```


Spotify

Hello. Hello hello and welcome to v. No nothing the podcast where I probably know just as much as you but with a glass of wine in my hand. I know even less. Maybe way less. I am Lucy from Lucy on the ground the mean / pop culture Instagram account that talks about everything from what the hell Kanye is doing with this church to why Meg Ryan is a rom-com goddess. Yeah, we're gonna get into that tonight with me against his will once again is the one and only Bill welcome Bill hello to your living room. Thank you for last week. You actually promoted me. Your Facebook and also begged anyone to replace you as the guests. I'm also your booking agent now to yeah, it's a few just taking all the I didn't hire you for that. Yeah, how's it going? I had a couple of responses. You might have some quality acts lined up coming up. I feel like I saw some of those responses and some of those are not didn't say they were all quality. I just had a couple quotes. Okay. Well, we'll see you guys but a lot of you have...


so how are
llections?
unfinished
results preview
contributions sought!

some descriptive statistics, geared towards my personal interests

Podcasts Vary in Content

Varying levels of professionalism

episodes professionally produced, by trained presenters

episodes from amateurs using a podcast-creation app (low barrier to entry)

Varying categories

Varying Audio Attributes

Background music, ads,
People speaking over each other

Languages

English, Spanish, French, Mandarin, Russian, Indonesian, Hindi, German, multilingual, code switching, ...

Spotify

What might we believe the differences between speech and writing are?

- Fleeting
  - speech is made and used in the moment and is not saved and … l
- Personal
  - speech is used in situations wh… present
- Elaboration
  - sp… g; writing needs explicitness to be
    … speech can rely on shared understanding
- …ion
  - speech is stuck to the conversational situation; writing is general and can have unbound variables
- Use cases
  - speech is used for different purposes than writing is

This is no longer (as) true

e.g Ong (1982) *Orality and Literacy* or Biber (1991) Variation across speech and writing or
Bakthin (1979) Проблема речевых жанров

Spotify

... but the conventional differences remain and are interesting to study.

and we can expect them to change as the genres and formats evolve in the near future.

What might we believe the differences between speech and writing are? (More concretion)

- Anchoring
  - the language is used in the situation and meant to be understood right then
- Subjectivity
  - the speaker is more present in the speech situation than the author is in the reading situation
- Discourse management
  - speech is less planned than writing and there will be explicit conversational moves related to the organisation of the discourse

compared to the Brown corpus

Feature	Podcast transcripts	Brown corpus
1st person singular pronouns	4.3%	0.19% (Press, reviews) - 2.3% (Romance novels)
1st person plural pronouns	1.3%	0.21% (Press, reviews) - 0.97% (Religious texts)
2nd person pronouns	3.6%	0.038% (Research) - 11% (Romance novels)
Amplifiers	0.71	0.15% (Press, reportage) - 0.35% (Press, reviews)

Francis and Kucera. 1967. Computational analysis of present-day American English. Providence. Brown University Press.

Spotify

Anchoring

- referential expressions
  - pronouns
  - demonstratives
- here-and-nowness
  - here
  - now
- tense-mood-aspect
  - reporting tense in written language is conventionally past
  - in spoken language this may not hold

Subjectivity

- attitudinal language
- amplifiers and hedges
- verbs of utterance
- verbs o Not done yet n
- verbs of perception

Discourse management

- cues
- interruptions
- repetition
- repair

Not done yet

Spotify

Comparison

AP news (from TREC data set)

Twitter (2017, Harvey storm)

Blogs (Authorship corpus)

Podcast data set

Switchboard corpus of transcribed telephone conversations

Movie scripts from UCSC

Spotify

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
Number of sentences	100 000	100 000	100 000	100 000	100 000	100 000
Number of words	2 400 000	1 800 000	2 100 000	1 900 000	120 000	890 000
Year of publication	1989-1990	2004	2017	2019	before 2010	early 1990s

Spotify

podcasts

... Only on my hands no with my hips ever. So first what I did was visiting a doctor because every time when I was trying to stretch myself like to take stretch classes, I ended up with like a really bad pain for like a few weeks or months. So then they visited doctor and I really like he told me that my spine like ...

movie scripts

- What's that shit?

- A book. It's called reading. You should try it some time.

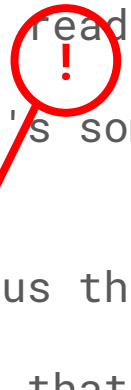
- You wanna read something. Read between the lines.

- Well here's something even you can relate to. Albert got a lotta trim.

- That genius thing is a babe magnet.

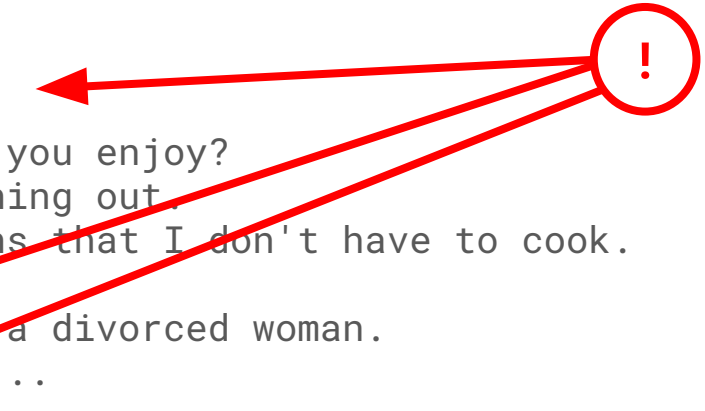
- Lemme see that book.

...



Spotify

phone conversations



--- What kind of ...
--- Okay.
--- ... eating out do you enjoy?
--- Well, I like dining out.
--- Of course, it means that I don't have to cook.
--- Right .
--- But, um, I'm a divorced woman.
--- I have one child ...
--- Uh-huh.
--- ... and, you know, when, when we dine out we go to like medium priced restaurants.

Spotify

Table 2: Occurrence of evaluative items.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
Positive	39 000	43 000	22 000	46 000	2 200	17 000
Negative	57 000	39 000	41 000	29 000	2 500	9 000

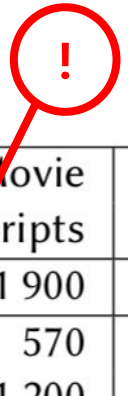
Table 3: Amplifiers.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
amplifiers	3 500	8 000	1 800	13 000	340	5 200
gradation	2 100	3 100	1 100	4 300	180	1 500
affirmation	730	4 300	280	7 500	120	3 500
surprise	640	680	510	970	28	212

Spotify

Table 4: Negation.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
negations	17 000	24 000	6 800	28 000	1 900	12 000
”no”, ”not”	11 000	10 000	3 300	11 000	570	4 100
contractions	4 000	12 000	1 800	15 000	1 200	7 500
constructions	2 000	2 600	1 700	2 000	160	790



Spotify

Table 5: Interjections and profanities.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
Interjections	~100	6 400	300	8 100	260	18 000
Profanities	~0	2 900	740	2 700	350	42

Table 6: Questions.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
Questions	720	7 100	2 700	7 600	2 400	3 700

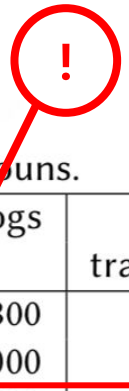


Table 7: Personal pronouns.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
1 person singular	8 000	95 000	5 300	77 000	4 400	36 000
2 person	2 400	16 000	7 000	52 000	3 000	18 000
3 person singular masculine	22 000	13 000	3 600	14 000	2 600	33 000
3 person singular feminine	4 400	8 200	1 600	6 400	600	1 900
1 person plural	5 300	12 000	6 600	18 000	730	7 700
3 person plural	11 000	7 000	5 200	12 000	340	9 700

Spotify

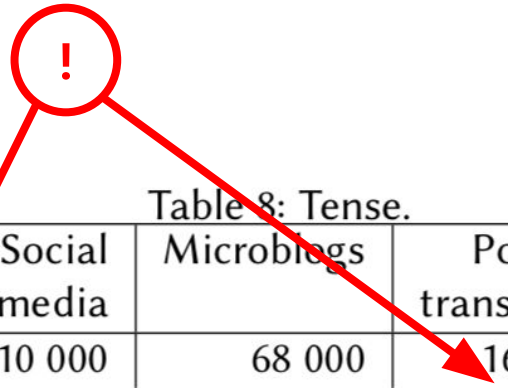


Table 8: Tense.

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
present tense	74 000	110 000	68 000	160 000	8 900	73 000
past tense	120 000	65 000	34 000	56 000	2 600	20 000

for training language models,
benchmarking is important

topicality is at the fore in benchmarks

Test sets	Size	Nouns or NP	Adjectives	Verbs	Other	Reference	Year
RG	65	100%				[15]	1965
Chiarello et al	144	100%				[3]	1990
TOEFL	80	21%	25%	21%	32%	[11]	1997
WordSimilarity	353	97%	1%	2%		[6]	2001
ConceptSim		100%				[16]	2011
BLESS	200	100%				[2]	2011
Entailment	15 992	100%				[1]	2012
Syntactic Analogy	8 000	25%	37.5%	37.5%		[12]	2013
SIMLEX	999	67%	11%	22%		[8]	2016

Karlgren, Jussi. 2019. "How lexical gold standards have effects on the usefulness of text analysis tools for digital scholarship." In *Proceedings from theInternational Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*.

Spotify

and if the character of podcasts is dramatically different from other data, this will have effects

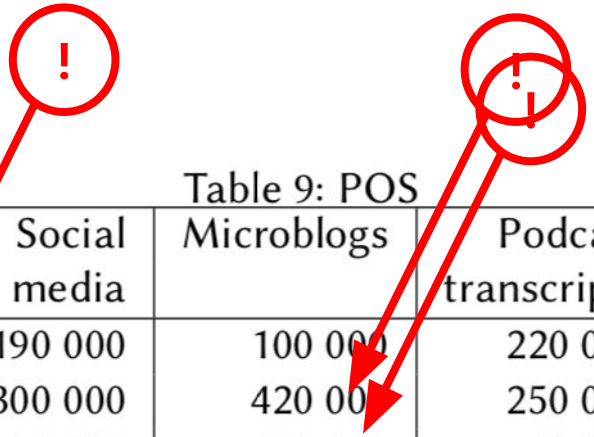


Table 9: POS

	Editorial media	Social media	Microblogs	Podcast transcripts	Movie scripts	Phone conversations
all verbs	190 000	190 000	100 000	220 000	11 000	93 000
nouns	470 000	300 000	420 000	250 000	16 000	100 000
proper nouns	260 000	92 000	490 000	63 000	7 200	20 000

Spotify

these were all early first scratches at the data set and intended to inspire more sophisticated study of the data

TREC Podcasts Track

Spotify

Task 1: segment search in podcasts

Spotify

given a topic, find 120s segments
from the podcasts that are relevant

Topics and Queries

in traditional TREC style

Terse Query
and
Slightly Wordier Description

three flavours

```
<topic>
<num>56</num>
<query>gaslighting</query>
<type>known item</type>
<description>On Twitter I saw someone reference a podcast interview with a doctor about gaslighting within organizational systems and I would like to hear it.
</description>
</topic>
```

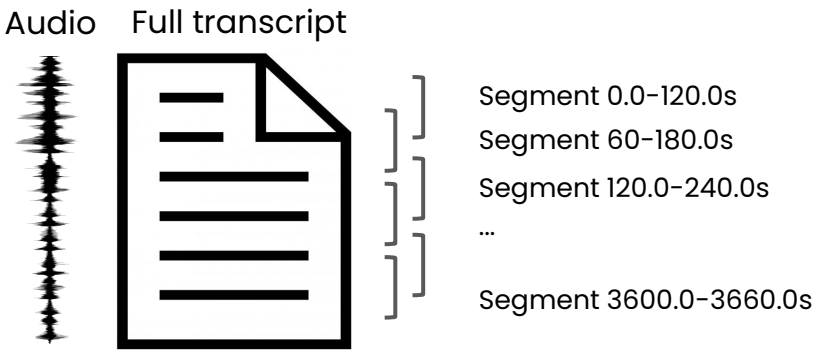
```
<topic>
<num>41</num>
<query>gmo food labeling</query>
<type>topical</type>
<description>
Some people say we should avoid foods with genetically-modified organisms as ingredients. I would like to learn about GMO food labeling. What are people saying about the pros and cons? What's the difference between the European and US approaches to GMO food labeling?
</description>
</topic>
```

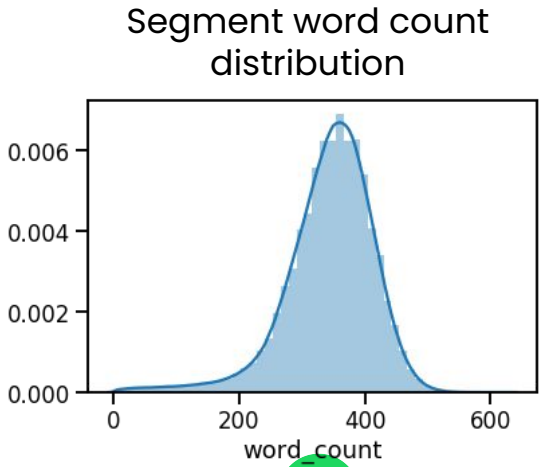
```
<topic>
<num>47</num>
<query>sci-fi author interview mars</query>
<type>refinding</type>
<description>I heard this interview with a sci-fi author who wrote a book or books with "Mars" in the title, but I can't recall his name.  I'd like to find it again.
</description>
</topic>
```

Spotify

Segments:

- 3.4 M podcast segments with 120s sliding window & 60s overlap
- word counts: 340 ± 70





Spotify

Scoring

Manual human assessment by NIST assessors on a 5-point PEGFB scale:

- **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)** A highly satisfying segment. Highly relevant to the information need / description, fully or mostly on topic, good entry point into the topic within the episode.
- **Good (2)** A somewhat satisfying segment. Relevant to the information need / description, mostly on topic, decent entry point into the topic within the episode.
- **Fair (1)** An unsatisfying segment. Limited relevance, only partially on topic, poor entry point into the topic.
- **Bad (0)** No relevant text in the segment.

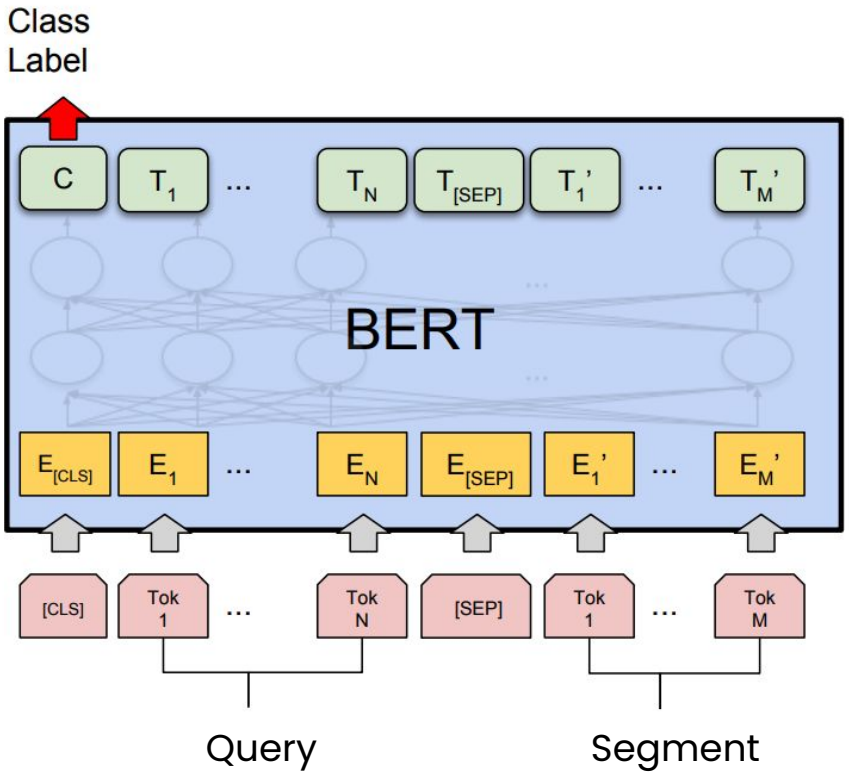
Spotify

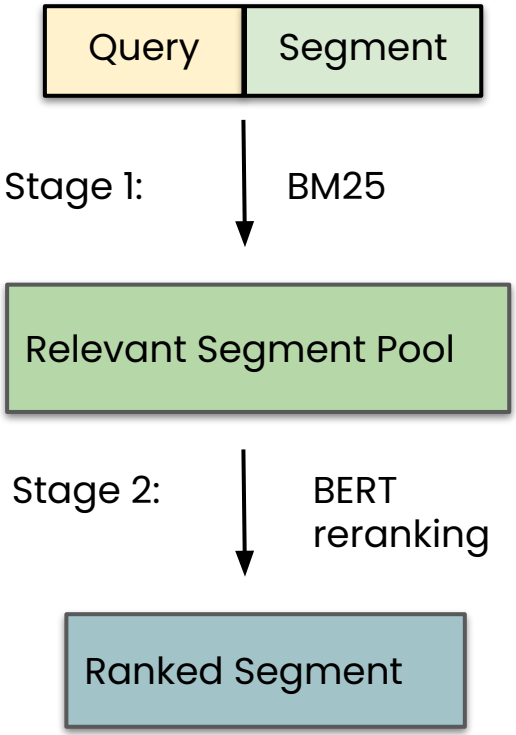
Evaluation Metrics

- Normalized Discounted cumulative gain (nDCG)
- Normalized Discounted cumulative gain at top of list (nDCG at 30)
- Precision at 10

Spotify

Baseline Models: BERT reranking





$$L = -\sum_{j \in J_{\text{pos}}} \log(s_j) - \sum_{j \in J_{\text{neg}}} \log(1 - s_j),$$

Spotify

results

	nDCG
UMD_IR_run3	0.67
UMD_ID_run4	0.66
UMD_IR_run1	0.62
UMD_IR_run5	0.65
UMD_IR_run2	0.59
run_dcu5	0.59
run_dcu4	0.58
run_dcu1	0.57
run_dcu3	0.57
run_dcu2	0.55
LRGREtvrs-r_2.	0.54
LRGREtvrs-r_1.	0.54
LRGREtvrs-r_3.	0.51
hltcoe4	0.51
hltcoe3	0.5
hltcoe2	0.47
hltcoe1	0.45
BERT-DESC-S	0.43
BERT-DESC-TD	0.43
BERT-DESC-Q	0.41
hltcoe5	0.38
UTDThesis_Run1	0.34
oudalab1	0

BM, QL
0.52

RERANK-Q
RERANK-D
0.43

	nDCG@30
UMD_IR_run3	0.52
UMD_IR_run5	0.5
UMD_ID_run	0.49
BERT-DESC	0.47
BERT-DESC	0.47
UMD_IR_run1	0.45
BERT-DESC-Q	0.45
run_dcu5	0.43
hltcoe4	0.43
UMD_IR_run2	0.42
run_dcu4	0.42
run_dcu1	0.42
run_dcu3	0.42
run_dcu2	0.4
hltcoe2	0.38
hltcoe3	0.35
UTDThesis_	0.34
hltcoe1	0.33
hltcoe5	0.3
LRGREtvrs-r_3.	0.32
LRGREtvrs-r_2.	0.40
oudalab1	0.40
LRGREtvrs-r_1.	0

RERANK-D
RERANK-Q
0.48-0.47

BM, QL
0.4

	p@10
UMD_IR_run3	0.6
UMD_IR_run5	0.58
BERT-DESC-S	0.57
UMD_ID_run4	0.56
BERT-DESC-TD	0.56
run_dcu5	0.54
run_dcu4	0.54
hltcoe4	0.54
UMD_IR_run1	0.53
BERT-DESC-Q	0.53
UMD_IR_run2	0.51
run_dcu1	0.5
run_dcu3	0.5
run_dcu2	0.48
LRGREtvrs-r_2.	0.48
LRGREtvrs-r_1.	0.47
hltcoe2	0.45
hltcoe3	0.43
UTDThesis_Run1	0.43
LRGREtvrs-r_3.	0.41
hltcoe1	0.38
hltcoe5	0.37
oudalab1	0.01

RERANK-D
RERANK-Q
0.57-0.56

BM, QL
0.49-0-48

this task was approachable without using the audio data

we would like to change that!

(but change as little as possible)

New Topic Types Coming Up!

```
<topic>
  <num>2</num>
  <query>...</query>
  <type>opinion</type>
  <description></description>
</topic>
<topic>
  <num>4</num>
  <query>...</query>
  <type>discussion</type>
  <description></description>
</topic>
<topic>
  <num>5</num>
  <query>daniel ek interview</query>
  <type>entertaining</type>
  <description></description>
</topic>
```

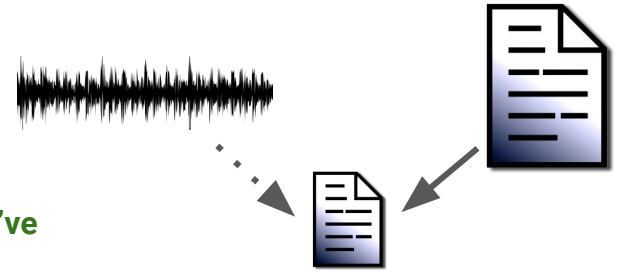
Spotify

Task 2: summarisation

Spotify

Podcast Summarization

Hello. Hello hello and welcome to v. No nothing the podcast where I probably know just as much as you but with a glass of wine in my hand. I know even less. Maybe way less. I am Lucy from Lucy on the ground the mean / pop culture Instagram account that talks about everything from what the hell Kanye is doing with this church to why Meg Ryan is a rom-com goddess. Yeah, we're gonna get into that tonight with me against his will once again is the one and only Bill welcome Bill hello to your living room. Thank you for last week. You actually promoted me. Your Facebook and also begged anyone to replace you as the guests. I'm also your booking agent now to yeah, it's a few just taking all the I didn't hire you for that. Yeah, how's it going? I had a couple of responses. You might have some quality acts lined up coming up. I feel like I saw some of those responses and some of those are not didn't say they were all quality. I just had a couple quotes. Okay. Well, we'll see you guys but a lot of you have...



This week Bill and I talk about our weekend drinking habits, what we've been up to, and why Meg Ryan is a rom com goddess.

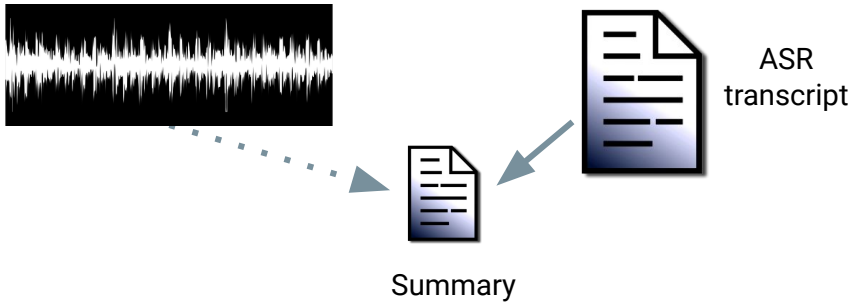
Summary

Task: Could you decide whether to listen based on this summary?

Spotify

The Podcast Summarization Task

Given a podcast episode with its audio and transcription, return a short text snippet capturing the most important information in the content. Returned summaries should be grammatical standalone passages of significantly shorter length than the input episode transcript.



Spotify

a gold standard is not that easy to come by!

Spotify

PODCAST EPISODE

How to Stop A Killer Asteroid

Science Vs

Dec 2019 · 31 min

Episode Description

This week — asteroids. Could a space rock really slam into us and destroy the world? And if we did spot one heading straight for us, is there anything we could do to stop it? We speak with asteroid researcher Dr. Alan Harris, astrophysicist Dr. Sergey Zamozdra, computational physicist Dr. Cathy Plesko, and physicist Dr. Andy Cheng.

Check out the full transcript here: http://bit.ly/2MrW1vp

Selected references:

Overview of Chelyabinsk impact and risk from asteroids: http://bit.ly/2ECSRQQ

How many asteroids are out there? http://bit.ly/34EhyHl

DART mission overview: http://bit.ly/2SkBBZ1

Ways to stop asteroids: https://bit.ly/2sJqGgv

Credits: This episode was produced by Wendy Zukerman along with Lexi Krupp with help from Michelle Dang, Meryl Horn and Rose Rimler. We're edited by Caitlin Kenney. Fact checking by Michelle Harris. Mix and sound design by Peter Leonard. Music written by Peter Leonard, Bobby Lord and Emma Munger. Recording assistance from Verónica Zaragovia, Sofi LaLonde, Lawrence Lanahan, and Kevin Caners. Translation help from Andrew Urodov and Dmitriy Tuchin. Thanks to all the scientists we spoke to: Dr. Carrie Nugent, Dr. Mark Boslough, Dr. David Kring, Dr. Daniel Durda, Dr. Kelly Fast and the other Dr. Alan Harris. A big thanks to Carl Smith at The Australian Broadcasting Corporation for suggesting this topic - Carl did a podcast series on a bunch of the Apocalypse scenarios! You can find it at the podcast Science Friction and search for the Apocalypse series. And thanks to the Zukerman Family and Joseph Lavelle Wilson.

show less

PODCAST EPISODE

JOSH GORDON IS BACK EMERGENCY PODCAST!

Title Talk

Aug 2019 · 30 min

Episode Description

JOSHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
GORDOOOOOOOOOOOOOOOOOOOOOOOOOOOONNNNNNNNNNNNNNNNNNNN!

This episode is sponsored by

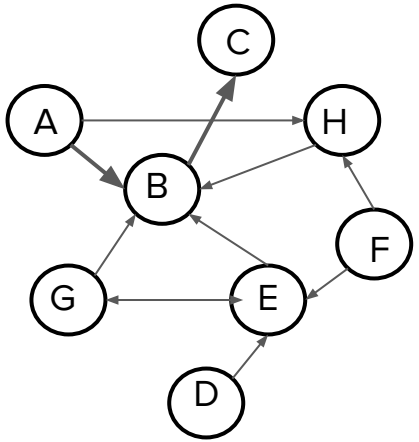
· Anchor: The easiest way to make a podcast.

https://anchor.fm/app

Support this podcast: https://anchor.fm/TitleTalk/support
show less

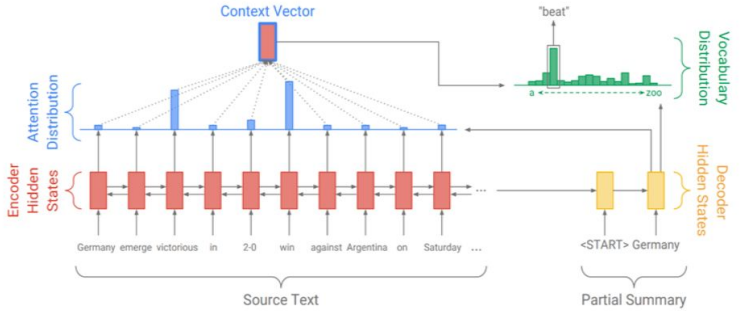
Spotify

Two Kinds of Approaches



Extractive Summarisation:

Identify and select phrases or sentences within the podcast that contain the most salient content
[eg. TextRank, LexRank]



Abstractive Summarisation:

Generate new summary content
eg. Neural generative summarisation

Spotify

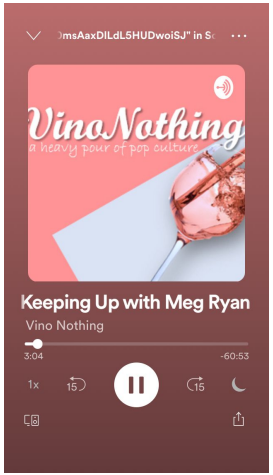
Sample Summaries

TextRank

All of the classics and a lot of you I've spoken to some of you and it was a little bit of well, it's not technically a rom-com or it's not all happy and I just know it's no it's not I don't need Insight from you because a lot of people are also throwing shade at you. Yeah, I mean 10 Things I Hate About You is a I wouldn't be upset with either one just because I I hope that it would lose to Love Actually in the final four and that it's going to be clueless and pretty woman in the final four and God only knows what you all have to I honestly I have a feeling coulis is winning this whole thing.

BART-Podcasts

This week Bill and I talk about our weekend drinking habits, what we've been up to, and why Meg Ryan is a rom com goddess. --- This episode is sponsored by · Anchor: The easiest way to make a podcast.Share



Spotify

Scoring

Manual human assessment by NIST assessors on a 4-point PEGFB scale:

- **Excellent (3)**
  Accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. It contains almost no redundant material which isn't needed when deciding whether to listen.
- **Good (2)**
  Conveys most of the important attributes and gives the reader a reasonable sense of what the episode contains. Does not need to be fully coherent or well edited. It contains little redundant material which isn't needed when deciding whether to listen.
- **Fair (1)**
  Conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain some redundant material which isn't needed when deciding whether to listen.
- **Bad (0)**
  Does not convey any of the most important content items of the episode or gives the reader an incorrect sense of what the episode contains. It may contain a lot of redundant information that isn't needed when deciding whether to listen to the episode.

System Quality: aggregated weighted average of episode scores, with E=4, G=2, F=1, B=0

Spotify

Baselines

Systems May Do Better than Creator-Generated Summaries

	E	G	F	B	Quality
Creator Descriptions	0.1564	0.2402	0.3408	0.2626	1.45
Cleaned Creator Descriptions	! 4	0.2123	0.3296	0.2737	**1.51**
First One Minute	0.0279	0.1397	**0.5363**	0.2961	0.93
TextRank on Sentences	0.0056	0.0168	0.1732	**0.8045**	0.23
TextRank on Segments	0.0168	0.033[illegible]	0.2458	0.7039	0.38
BART Pre-trained on News	0.0559	0.13[illegible]	0.4916	0.3128	0.99
BART Fine-tuned on Training	0.1397	**0.2793**	0.3743	0.2067	**1.49**

Spotify

TREC 2021 coming up!

- sign up! we are working towards lowering the participation threshold for new entrants!

Spotify

Use the corpus for other study!

The Podcasts Corpus is relevant to all your favourite empirical linguistic questions!

- Sociolinguistics
- Speech processing
- Acoustic aspects of the domain
- Stylistics
- ...

very much a team effort

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich (CLARIN EU), Gareth J.F. Jones (Dublin City U), Jussi Karlgren, Aasish Pappu, Sravana Reddy, Yongze Yu

plus all the participating teams at TREC!

Cambridge U, Johns Hopkins, Uppsala U, U Cambridge, U Central Florida, U Delaware, U Glasgow, U New Hampshire, U Texas Dallas, Dublin City U, U Maryland, U Texas Dallas, U Oklahoma, Birla Institute of Technology and Science

Spotify