CC-News-En: A Large English News Corpus

Joel Mackenzie
The University of Melbourne
Melbourne, Australia

Rodger Benham
RMIT University
Melbourne, Australia

Matthias Petri
Amazon Alexa
Manhattan Beach, CA, USA

Johanne R. Trippas
The University of Melbourne
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

ABSTRACT

We describe a static, open-access news corpus using data from the Common Crawl Foundation, who provide free, publicly available web archives, including a continuous crawl of international news articles published in multiple languages. Our derived corpus, CC-News-En, contains 44 million English documents collected between September 2016 and March 2018. The collection is comparable in size with the number of documents typically found in a single shard of a large-scale, distributed search engine, and is four times larger than the news collections previously used in offline information retrieval experiments. To complement the corpus, 173 topics were curated using titles from Reddit threads, forming a temporally representative sampling of relevant news topics over the 583 day collection window. Information needs were then generated using automatic summarization tools to produce textual and audio representations, and used to elicit query variations from crowdworkers, with a total of 10,437 queries collected against the 173 topics. Of these, 10,089 include key-stroke level instrumentation that captures the timings of character insertions and deletions made by the workers while typing their queries. These new resources support a wide variety of experiments, including large-scale efficiency exercises and query auto-completion synthesis, with scope for future addition of relevance judgments to support offline effectiveness experiments and hence batch evaluation campaigns.

ACM Reference Format:
Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A Large English News Corpus. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412762

1 INTRODUCTION

Reproducible and replicable experimentation is an indispensable component of research. In the field of information retrieval (IR), enormous effort has been expended on reproducibility best practices, including various challenges such as the Open Source Information Retrieval Replicability Challenge (OSIRRC) [19, 29]; dedicated reproducibility tasks based on NTCIR, CLEF, and TREC (CENTRE) [22]; and, since 2015, a dedicated reproducibility track at ECIR. Indeed, conferences such as NTCIR, CLEF, and TREC routinely host a number of *tracks* that produce data such as queries and relevance judgments over different target corpora.

One problem, however, is that many commonly used document corpora are not freely available, which can impede reproducibility efforts for groups without access to these resources. Thankfully, a range of initiatives mean that large-scale open-source data is becoming more readily accessible. One such example is the Common Crawl Foundation,[1] who generate large-scale crawls of the web at regular intervals. A key philosophy behind the Common Crawl is to *democratize data*, allowing open access with no fees. In late 2016, the Common Crawl Foundation announced a *news-specific* crawl (CC-News),[2] with documents being added on a *daily basis*, and covering sources from a wide range of countries and languages.

Here we derive a static, English segment of the CC-News crawl that we refer to as CC-News-En. Due to the storage and computation costs involved in filtering out non-English documents, we make the complete corpus available as a free resource, along with a suite of tools which can be used to replicate corpus extraction from the original source CC-News data. We also provide a set of 10,437 *user query variations* over 173 query topics, including keystroke-level data collected from a novel crowdworking experiment. Our goal is to encourage reproducible and replicable experimentation, with greatly reduced barriers to entry.

Contributions. This project results in four key contributions:

(1) A large, freely available, English news collection based on the Common Crawl news corpus;
(2) A set of crowdsourced user query variations which correspond to news events contained within the corpus;
(3) Matching keystroke query-entry data from crowdworkers; and
(4) A range of tools which can be employed to replicate, analyze, and extend the document corpus, and to assist with creating data for crowdsourcing experiments.

Section 2 provides the background motivation for CC-News-En. Section 3 outlines how CC-News-En is composed and contrasts it with other text corpora. Section 4 describes the novel approach used to develop topics, along with the crowdsourcing study used to solicit query variations. Section 5 explores the practical utility of

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412762

[1] https://commoncrawl.org/
[2] https://commoncrawl.org/2016/10/news-dataset-available/

CC-News-En in the context of past studies, and lists some potential limitations and lessons learned. Finally, future work is outlined and a summary of the new collection is provided in Section 6.

2 BACKGROUND AND MOTIVATION

Text Collections and News Retrieval. Construction of reusable test collections for repeatable experimentation is a major activity in the field of IR. Typical test collections follow the Cranfield paradigm [20], and comprise a set of documents, a set of topics, and a set of associated relevance judgments [44]. Most test collection usage occurs in the academic community through conferences and workshops such as TREC, NTCIR, and CLEF. News corpora have a been mainstay in such experimentation, with many of the early TREC campaigns making use of full-text newswire articles [44]. The main flavor of such tasks was ad-hoc retrieval, using news corpora typically containing a few thousand to a few hundred thousand documents, as provided by large news organizations. These documents were often written by professional journalists and subject to careful copy editing, making them a high-quality data resource.

News retrieval tasks continue to be important. For example, the "NewsIR" workshop [2] focuses on the "possibilities and challenges that technology offers to the journalists, the challenges that new developments in journalism create for IR researchers, and the complexity of information access tasks for news readers".[3] Furthermore, TREC is currently hosting a dedicated news track [45] which focuses on the *"news user"*, with tasks such as "background linking", recommending the next articles a user should read to contextualize the query article; and "entity ranking", providing links to relevant entities within the query article to support the understanding of the reader. Recently, a number of studies have used variations (due to its dynamic nature) of the CC-News corpus for efficiency experimentation, including studies on index re-ordering [34] and index compression [42, 43].

Crowdsourcing and Query Generation. Alonso et al. [3] were the first to note that expert relevance assessors might be replaced by Amazon Mechanical Turk crowdworkers to reduce assessment costs, and to explore the cost–quality trade-off between crowdworkers and expert assessors. Alonso et al. also observe that the crowdworker may be completing an artificial task (that is, a hypothetical task) and that tasks could be culturally-specific. Checco et al. [16] give consideration to different attacks that crowdworkers might employ (human or otherwise) to pass the *gold questions* used as quality control in some studies. Despite these quality concerns, which can be hedged to varying degrees but never fully eliminated, crowdsourcing is now a popular technique that allows practitioners to carry out IR user studies faster and with less expense than previously [6, 24, 33, 39].

Bailey et al. [8] also employed crowdsourcing, constructing a test collection that associates multiple user queries with each information need, with each of those needs expressed as a personalized text *backstory* derived from a single TREC topic. Having a set of *user query variations* associated with each of the TREC topics, rather than just a single query, has enabled enhanced understanding in a range of areas: test collection judgment pool methodology [37]; the consistency [9] and risk [11] properties of retrieval models; the quality of automatic query generation approaches [32]; and new implementation options for efficient search on web corpora [14]. Similar query collections have also been created by teams working on TREC-initiated activities [12, 13], adopting the notion of an information need expressed as a backstory, and also adopting the previous mode of presentation of the backstory, as text to be read by the crowdworker. One of the goals of our new collection is to vary how backstories are *presented*, to further diversify the pool of query variations available for subsequent study.

Audio Transcriptions in IR. Images have been used to solicit query variations in medical IR [47]. Similarly, the rise in popularity of spoken-text retrieval devices means that studying how searchers form queries after listening to an audio snippet will be useful. Spina et al. [46] show that query-biased document summaries presented as audio are practical in conversational IR. Providing snippets through speech synthesis introduces more presentation factors, as Chuklin et al. [18] note, where read-outs with prosody changes were subjectively more informative, at the expense of their aesthetic quality. We explore audio read-outs of backstories and include the user query variations that were generated via that modality as part of the new CC-News-En collection.

Known-Item Search. Azzopardi and de Rijke [5] compare real and automatically generated query variations for a known document in a collection by independently drawing query terms. They note that *"this assumption eliminates the need for explicit relevance judgments as the known-item is the relevant document"*. We adapt this approach to gather real user queries to make available as part of this reproducible corpus. Known-item search was also used in the TREC-7 spoken document retrieval track [23].

3 CORPUS

This section describes the CC-News-En corpus and how it was built from the original CC-News data, including an analysis of the characteristics of the corpus.

3.1 Construction

Common Crawl Data. The starting point for CC-News-En is the data from Common Crawl, which we refer to as CC-News. This data is crawled using a variation of StormCrawler,[4] which itself is based on Apache Storm. Each day, a new set of WARC files is added to CC-News, and hosted on AWS S3. Further details are provided in the Common Crawl news-crawl[5] repository.

Commencement Set. A total of 2,291 CC-News WARC files were processed to build CC-News-En, covering the period 26 August 2016 to 31 March 2018, inclusive. The first and last WARC files in this collection are as follows:

- `CC-NEWS-20160826124520-00000.warc.gz`
- `CC-NEWS-20180331191315-00143.warc.gz`

The resulting subset of compressed WARC files occupies 2.14 TiB of disk space, and contains a total of 102.5 *million* documents in over 100 languages.

[3] https://research.signal-ai.com/newsir19/

[4] http://stormcrawler.net/

[5] https://github.com/commoncrawl/news-crawl

Filtering English Documents. Since the CC-News data is not constrained to any particular language, a large fraction of the commencement set is in languages other than English. To remove documents that were not English, the entire collection of WARC files was filtered as follows. Firstly, each WARC file was read into memory and decompressed. Secondly, each document in it was parsed using Apache Tika[6] and the underlying text was analyzed with Apache OpenNLP.[7] If the document was predicted to be English, it was retained; otherwise, it was discarded. Finally, the retained documents were written back to disk in a modified WARC file. A similar methodology has been used in previous experiments on the Common Crawl [38, 41]. The Java program took close to two days to filter and re-write all 2,291 WARC files, with processing performed on a Linux machine with two Intel Xeon Gold 6144 CPUs at 3.50GHz and 512 GiB of RAM, with all I/O operations via a local RAID array, using 32 worker threads.

Augmenting TREC Identifiers. A desirable property of our target collection is to have a unique, human-readable identifier for each document. Following the approach of the Lemur Project,[8] a custom field was added to each document's WARC response header named `WARC-TREC-ID`. This identifier describes the location of the document within the compressed collection, and is formatted as:

`CC-NEWS-⟨timestamp⟩-⟨serialno⟩-⟨record⟩`,

where `timestamp` refers to the timestamp in which the WARC was created in the original CC-News corpus, `serialno` refers to the crawl process identifier from the original corpus, and `record` is the k-th document seen within the WARC file, assigned during language filtering. For example, the identifier

`WARC-TREC-ID: CC-NEWS-20180214191955-00245-2`.

corresponds to the second document from the unfiltered

`CC-NEWS-20180214191955-00245.warc.gz`

file. Note that the `record` value is assigned prior to filtering, and may not be consecutive in the filtered CC-News-En corpus – two contiguous documents from the same WARC file with `record` numbers of 5 and 8 indicate that the 6 th and 7 th documents were analyzed as non-English, and were removed.

Final Corpus. After the filtering process, 2,290 WARC files remained, one fewer than the original commencement set (see Section 5 for more information). Table 1 summarizes the gross attributes of the CC-News-En collection, which is freely available from AARNet's CloudStor platform via the URL provided in Table 3, under the same terms of use as is the original CC-News.[9] Additionally, we provide a CC-News-En *Common Index File Format* [31] index derived from the Anserini [49] toolkit for improved usability.

3.2 Characteristics

To appreciate the nature of CC-News-En, we provide an analysis of its characteristics, with Indri 5.11 used to generate a Krovetz-stemmed index. Note that a different indexing pipeline might result in slight differences to the statistics reported here.

[6]https://tika.apache.org/1.13/
[7]https://opennlp.apache.org/docs/1.8.4/manual/opennlp.html
[8]http://www.lemurproject.org/clueweb09/datasetInformation.php
[9]https://commoncrawl.org/terms-of-use/, accessed 2 June 2020.

Table 1: Basic statistics for the raw CC-News-En corpus after filtering out non-English documents. The reported size refers to the compressed (gzipped) WARC files.

Size [GiB]	WARC Files	Total days	Date range
965.7	2,290	583	26/08/2016 – 31/03/2018

Table 2: The most popular news sites observed across CC-News-En. In total, there are close to 30,000 unique sources of news documents, with around 10,000 of those contributing just a single document.

Source Site	Pages	Percent	Type
`reuters.com`	7,707,626	17.7	News
`topix.com`	876,744	2.0	News, Forum
`dailymail.co.uk`	714,916	1.6	News
`ycombinator.com`	659,439	1.5	Aggregator, Forum
`cbslocal.com`	467,229	1.1	News
`einpresswire.com`	335,283	0.8	Press Release
`ohio.com`	302,860	0.7	News
`yahoo.com`	275,057	0.6	News
`cnn.com`	257,769	0.6	News
`bbc.{com,co.uk}`	254,692	0.6	News

Contributing Websites. Table 2 lists the ten most common sites observed within CC-News-En, and the proportion of documents that come from each. As expected, the most prominent sites in CC-News-En are those of large news companies such as *Reuters, Daily Mail, CBS,* and so on. There are also some news aggregators, and forums which reflect content from a number of sources, including comments from users of the respective sites. For example, *Topix* is an American website that originally focused on news aggregation, but moved into content creation and local news forums. Similarly, *Y Combinator* runs the high-traffic *HackerNews* site, which acts as both a tech news aggregator and also as a tech forum (via text posts and comment sections). Figure 1 presents the cumulative corpus size (by document count) as sources are added in decreasing order of popularity. The top ten sites (listed in Table 2) account for slightly more than 25% of the documents within the corpus, with a long tail of sources contributing just one or two documents each.

Temporal Growth. The Common Crawl foundation are constantly adding new documents to CC-News. In turn, this means that CC-News-En also captures this temporality. To illustrate this facet of CC-News-En, Figure 2 shows the number of documents added to the collection (left), and the cumulative size of the collection (right), grouped by whether the site was in the top 10 most popular or not. While the collection started slowly, the rate of growth increased rapidly in December 2016 as a result of the addition of DMOZ open directory data,[10] which increased the number of seed URLs for the news crawl.[11]

Comparison to Common IR Collections. Table 3 provides more context on the characteristics of CC-News-En relative to previous

[10]DMOZ is now superseded by Curlie: https://www.curlie.org
[11]https://github.com/commoncrawl/news-crawl/issues/8

Table 3: Statistics for commonly used document collections, after indexing via Indri with Krovetz stemming. The CC-News-En corpus is much larger than the Robust04, NYT, and Gigaword newswire corpora, and rivals the size of the Gov2 and ClueWeb12B web collections.

Corpus	Documents	Unique terms	Total terms	Total postings	Reference
Robust04	528,155	664,603	253,094,062	112,652,378	https://trec.nist.gov/data/cd45/index.html
NYT	1,855,658	2,970,013	1,277,892,472	501,568,918	https://catalog.ldc.upenn.edu/LDC2008T19
Gigaword	9,875,524	2,613,928	4,110,355,970	2,052,690,482	https://catalog.ldc.upenn.edu/LDC2011T07
Gov2	25,205,179	39,180,841	23,804,988,213	5,880,709,591	http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm
ClueWeb12B	52,343,021	165,309,502	39,795,590,329	15,319,871,265	http://lemurproject.org/clueweb12/
CC-News-En	43,530,315	43,844,574	49,789,013,621	20,150,335,440	http://go.unimelb.edu.au/u3nj

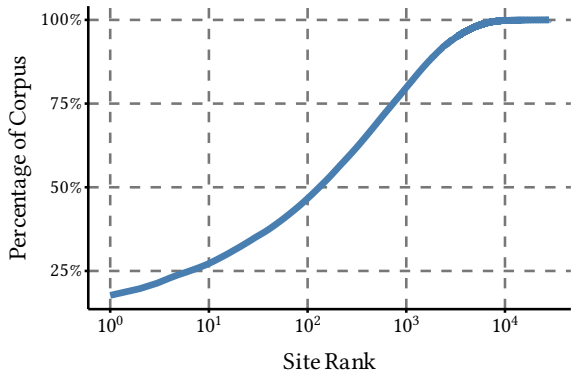

Figure 1: The cumulative percentage of the corpus, in terms of documents, as sites are added by decreasing popularity. The first ten largest sites account for over 25% of documents, and the top 100 sites account for close to 50% of all documents.

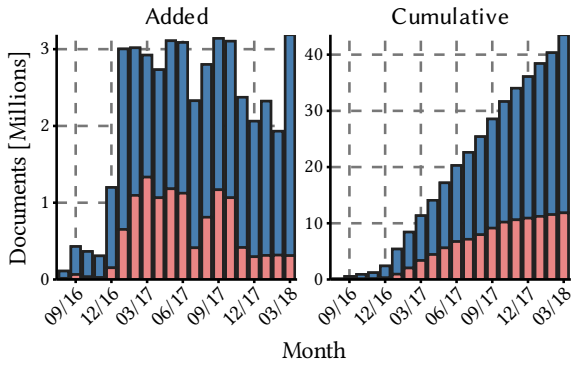

Figure 2: The number of documents added to the corpus (left) and the cumulative number of documents in the corpus (right), broken down by month. The light (bottom) portion of each bar represents documents from the top ten sites (listed in Table 2) and the dark (top) portion of each bar represents all other sites.

text collections, using the same Indri pipeline throughout. In particular, CC-News-En can be compared to the following:

- Robust04: Various newswire articles from the 1980s and 1990s;
- NYT: New York Times articles from 1987 to 2007;
- Gigaword: Various newswire articles from seven sources between 1994 and 2011;
- Gov2: .gov articles from early 2004; and
- ClueWeb12B: Web documents crawled in the first half of 2012.

Table 3 lists a number of basic statistics across these newswire and web corpora. As can be seen, CC-News-En is much larger than previous newswire resources, and is comparable in size to common web collections. While CC-News-En has fewer documents and unique terms than ClueWeb12B, it contains a greater number of total terms and postings. This suggests that CC-News-En has longer documents on average than a typical web corpus, and with the more concentrated vocabulary a consequence of news stories often using similar language and (perhaps) also containing fewer spelling errors.

4 TOPICS

While there are many open-source query logs available, these logs may not be topically or temporally relevant to the CC-News-En corpus. To improve the usability of the corpus, we also provide a large set of representative queries, created via crowdsourcing.

4.1 Desiderata and Basic Approach

To be suitable, a query log should contain queries pertaining to news stories that are both temporally relevant to, and covered by, documents in the CC-News-En corpus. To meet these desiderata, our approach generates summaries for a subset of documents from the corpus, and shows these summaries to crowdworkers, asking them to provide the query that they would issue if they wished to learn more about the topic. This approach simulates the backstory technique of Bailey et al. [7, 8], in which crowdworkers are shown a short information need statement, and asked what their first search query would be. A similar approach has been used to generate queries from community question answering sites [15].

As well as simply showing workers text statements and asking them to generate suitable queries, we also simulate a *breaking news* summary, similar to those heard over the radio or on TV, and ask the crowdworkers to listen to *audio* renditions of the same text statements. The goal in both elicitation modalities is to collect a diverse set of *query variations* for a non-trivial number of topics, which can serve as a query log for CC-News-En. Note that one beneficial side-effect of this methodology is that each of the summaries shown to crowdworkers is drawn from a document within the CC-News-En corpus, and hence also provides an associated implicit relevance signal. That is, while there is not (yet) a set of qrels for CC-News-En, there is nevertheless a sense in which retrieval effectiveness might also be assessed [5].

4.2 Collecting News Articles

Finding Temporally Relevant Target Articles. The first step in generating representative queries is to find a set of target news articles that temporally align with the CC-News-En collection. While we could simply generate a summary for *each* document in the corpus, this would likely result in many near-duplicate summaries since several sources typically report on each news story.

Instead, the Reddit API[12] was used to download the top 50 articles that were submitted from the popular `news` and `worldnews` subreddits for each of the 583 days spanned by CC-News-En. For each of the 58,300 Reddit threads identified by this process the URL, the title, the number of comments, and the number of associated upvotes were tabulated. The number of comments and upvotes given to each thread can be used to implicitly capture the popularity or impact of each news story. Furthermore, Reddit titles generally provide succinct and accurate summaries of news events, an important feature for validation of the approach.

Mapping Back to the Target Corpus. The second step involved mapping the data from Reddit *back* into the CC-News-En corpus, to ensure that each topic has at least one relevant item in the corpus. After joining the Reddit data with the CC-News-En corpus via pattern matching over URLs, just over 15,000 matches remained.

4.3 Generating Backstories

The next step is to generate summaries which can be shown to crowdworkers to solicit queries.

Summarization. Following recent work on news-related tasks [1, 50], we used the open-source Newspaper3k[13] tool to parse each target article and generate a summary of it. It implements extractive summarization, based on both term and positional features [25, 28]. We refer to this as the Extractive summarizer.

A basic "first-k sentences" summarizer [10] was also employed, given that the opening paragraph of news articles generally provides an accurate and succinct overview of the article [35, 40]. We refer to this option as the Intro summarizer. Recent work has shown that users can be sensitive to the length of summaries in both text and audio format, often finding longer summaries to be more informative [36, 48]. Thus, we employed three different summary lengths, thereby varying the extent of information provided to the crowdworkers, again with the goal of increasing the diversity of the user query variations:

- *Title*: The title of the article (may differ from the title submitted to Reddit);
- *Short*: A short summary consisting of the first sentence from the long summary; and
- *Long*: A long summary consisting of (up to) the top three ranked (or the first three) sentences from the article.

Since summarization was automated, we added an additional filtering stage to remove empty, excessively short, or excessively long summaries. In particular, we retained only titles containing between four and twenty words; short summaries containing between ten and fifty words; and long summaries containing between 50 and 100 words, inclusive. In planned future work we will explore how summary length may have influenced crowdworker behavior and the queries that were proposed.

[12] https://github.com/pushshift/api, accessed 2 June 2020.
[13] https://github.com/codelucas/newspaper, accessed 2 June 2020.

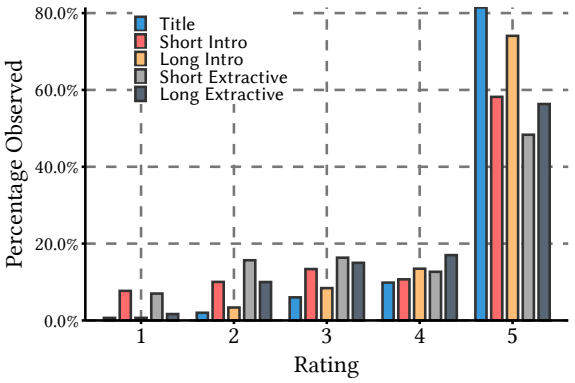

Figure 3: Ratings across the sampled topics. On average the introductory summaries were preferred to the extractive summaries.

Topic Curation. A curated subset of the topics that survived the filtering stages was then created. Fifteen viable topics from each month spanned by the collection were selected at random, and each was then inspected by two of a panel of six IR experts,[14] taking the Reddit thread title to be ground truth. In this blind experiment each expert considered a sequence of Reddit thread titles, document titles, and short and long summaries, with the latter two drawn from either the Extractive or Intro approaches at random; and for each of those summary options was asked to assess how accurately it conveyed the assumed intent of the Reddit title, using a five-point Likert scale, with five indicating "accurate". The experts were also asked whether they deemed the topic to be unsuitable for display to crowdworkers, such as ones discussing extreme violence or encouraging of other illegal behavior.

Figure 3 shows the results. Four of the five summary options had a median score of 5, the exception being the short Extractive summary, with a median score of 4. Document titles were most likely to express the intent of the Reddit title, a consequence of the Reddit titles often being very similar to the corresponding document titles. Of the two summarizer options, the Intro summaries were generally preferred, gaining higher average ratings than the Extractive summaries, for both short and long summary types.

For each topic the short and long summary options (Intro or Extractive) with the highest scores were then extracted, with ties broken by preferring the Intro-derived summary. At the same time, topics which did not have a rating of at least 4 across all of title, short summary, and long summary, were discarded, as were topics which were deemed ethically unsuitable. This process resulted in a set of 173 unique topics, 30 (17%) of which employed Extractive summaries, and 143 (83%) which employed Intro summaries.

4.4 Information Expression Formats

Two alternatives were used when presenting summaries to crowdworkers.

[14] The six authors of the current paper.

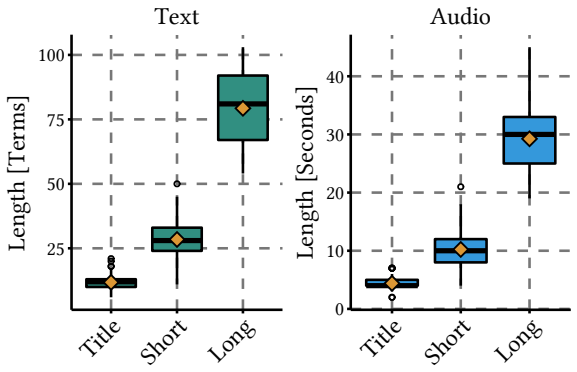

Figure 4: Summary lengths for textual summaries (left, in words) and the corresponding audio segments (right, in seconds). The diamond and horizontal line within each box represents the mean and median, respectively.

Image of Text. The first format involved written text, but presented as an image, to force the crowdworkers to fully type their queries (a pilot experiment with plain text having shown high rates of copy/paste from the summary statement). Disallowing copy and paste can be expected to lead to natural query generation, with increased use of synonyms, but also increased likelihood of spelling and typographical errors.

Spoken Summaries. The second format presented the summaries via spoken audio files. The Microsoft Cognitive Services *Speech Service* API was used to convert each summary into a speech file using the en-US-AriaNeural model, which outputs a female voice with a typical "United States" accent. To further simulate a news-based approach, the *newscast*[15] voice style was selected, which *"expresses a formal and professional tone for narrating news"*.

Figure 4 plots the length distributions of the corresponding textual and audio summaries, counting text in words, and speech in seconds. Titles have around 12 terms on average, with four-second audio clips. Short summaries are considerably longer, with an average of 28 terms and ten seconds per audio clip. Finally, the long summaries are 79 terms on average, and 29 seconds when read.

Instrumenting the Query Collection. The crowdsourcing interface collected keystroke information, information about the workers operating system, browser, and recorded device characteristics. This data was used to determine whether each worker used a mobile phone, tablet, or desktop/laptop computer, and whether they used a touch screen or keyboard for input. The interface also collected fine-grained statistics about worker input patterns and latency, and served as a protective mechanism to cull spam responses [17].

4.5 Query Collection

Process. The query collection was carried out with approval from the RMIT University ethics committee, and employed Amazon's Mechanical Turk[16] crowdsourcing platform. Workers were required to have completed at least 10,000 prior tasks, and have an acceptance rate of 95% or higher. Each HIT (human intelligence task) consisted of three tasks, all via the same modality; for example, a single HIT contained three spoken titles, or three long textual summaries.

Each topic has six unique ways of conveying the information need to the user; titles, short summaries, and long summaries, each being conveyed as either an image of text, or as an audio recording. Batches were deployed sequentially in a way that prevented workers from operating multiple modalities at once, to reduce familiarity bias [15]. Up to 25 workers were enlisted for each of the six formats.

Workers were paid according to the expected completion time of each task (measured in a pilot study) at an estimated wage of $USD7.50 per hour. Each completed batch was cleaned using a semi-automated approach to identify suspicious entries based on the time taken to complete a task, query similarity across topics, query length, and keystroke data. After identifying these suspicious entries and filtering false positives, a total of 135 HITs were rejected. Since there were workers who did most tasks correctly with only occasional suspicious tasks, the rejection decision was made on a per-HIT rather than a per-worker basis.

Table 4: Number of query variations for each topic, categorized by the format and length of the summaries shown to the crowdworkers.

	Text			Audio		
	Title	Short	Long	Title	Short	Long
Min	8	9	8	7	8	12
Mean	9.1	9.1	11.1	8.0	10.1	13.0
Max	19	18	20	16	20	25
Total	1,566	1,572	1,914	1,383	1,752	2,250

Query and Worker Characteristics. After rejecting suspicious HITs, a total of 195 unique workers contributed 10,437 query variations across the 173 topics, at an average of 53.5 queries per worker. Workers primarily used desktop or laptop computers to complete the HITs, with 8,978 (86%) of the queries submitted from either desktop or laptop computers, 985 (9%) submitted from mobile phones, and 126 (1%) submitted from tablets. The remaining 4% of queries were from unknown devices.

Table 4 shows the distribution of per-topic query variations, categorized by the format of the HIT. In total, each topic has between 56 and 118 associated variations, with an average of 60.3 per topic. As was also the case for the variations collected by Bailey et al. [8], each query was converted to lowercase, extraneous whitespace and punctuation was stripped, and the resulting query was passed through the Microsoft Cognitive Services *Bing Spell Check* API,[17] resulting in a final set of 9,947 unique normalized query variations, an average of 57.5 per topic. The normalized variations had a mean length of 7.09 terms, considerably longer than typical web search queries, but aligned with other recent work on crowdsourcing queries for complex information needs [15]. We plan to conduct an analysis of the worker/query/mode relationship in future work.

[15] https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/speech-synthesis-markup, accessed 2 June 2020.
[16] https://www.mturk.com/, accessed 2 June 2020.
[17] https://azure.microsoft.com/en-us/services/cognitive-services/spell-check/, accessed 2 June 2020.

5 DISCUSSION

We now discuss various use cases and limitations of the CC-News-En corpus, and some lessons learned from creating the collection.

5.1 Use Cases

Query variations can be leveraged to improve search engine efficiency and effectiveness; yet to date there has been no news (as distinct from web) collection that allows their study. Hence, of immediate interest is the facilitation of efficiency experimentation comparing retrieval models, driving new insights into scalable news retrieval. There are two facets to this: the new collection enables exploration of retrieval techniques on a static news collection as a repeatable baseline; plus the 10,437 user queries allow researchers to develop enriched query timing insights. Large IR collections such as Gov2 and the 2009 and 2012 ClueWeb crawls do exist, but they are focused on web rather than news content, and neither are freely available. As a publicly downloadable at-scale resource, CC-News-En thus provides unique opportunities, including the ability to explore the differences between web and news retrieval.

Another use-case of CC-News-En is the query-level keystroke information that it provides, which represents a new type of resource for academic researchers to explore. For example, the detail now available might lead to future work in understanding issues to do with query autocompletion [26, 27]; query reformulation; and (with suitable qrels added) query performance prediction. Replaying the CC-News-En queries to reproduce the way they were typed might also have educative value (and perhaps artistic value too), allowing IR practitioners to better empathize with the needs of searchers, and hence better tune retrieval models to match user behaviors. The queries might also help motivate digital literacy campaigns.

Finally, as each CC-News-En document has associated temporal data, it is possible to use it to simulate dynamic news streams, including tasks such as real-time news indexing, aggregation, and recommendation. Similar challenges such as the TREC temporal summarization [4] and real time summarization [30] tasks have been proposed in the past, and it could be interesting to explore these dynamic tasks through the lens of the CC-News-En corpus.

5.2 Limitations

Noise in the Corpus. As with any large document collection, there are many possible sources of noise. For example, news sites may occasionally place a link to an advertisement directly in their feeds, which might then be fetched by the crawler, resulting in non-news or even spam documents being incorporated into the document collection. Another possible issue is that of duplicate documents. Important news events often result in a number of near-identical articles appearing, containing similar entities, phrases, and term distributions, and making it difficult to determine whether documents are exact duplicates (the same article, perhaps re-published on a franchised site), near-duplicates (a different article covering the same story), or even totally different articles [21]. We did not attempt to de-duplicate the CC-News-En corpus, but note the challenge as a possible future research direction.

A further source of noise arises in the filtering process. With more than 50% of the original articles removed, some English documents may have been discarded (not necessarily a problem), and some non-English documents may have been included (more likely to be a problem). More robust methods of language identification might be worth exploring in future iterations of the collection.

Missing Documents and Temporal Gaps. During the creation of the collection, the `CC-NEWS-20170812163812-00038.warc.gz` file was not processed correctly by our pipeline, and was subsequently dropped from the CC-News-En corpus. In addition, there are six days within the 583 day period where no WARC files were added to the original CC-News crawl: 22/09/2016–25/09/2016 inclusive, 18/12/2017, and 22/12/2017. These gaps typically correspond to hardware and software upgrades on the crawl servers.[18] It is also important to note that both CC-News and CC-News-En are not intended to be complete crawls of their sources, but rather, to provide a reproducible sample of these sites.

Target Document Bias. Since the crowdworkers were asked to generate a query based on either a textual summary or an audio rendition derived from the document summary, the resultant queries may be biased towards keywords that appear in that document. As a result, the queries may be inadvertently focused on the target document from which the summary was created. We sought to reduce this bias and diversify the query pool by using three summary lengths (of varying informativeness) and two different modalities. In future work we will investigate *backstory variations*, where a number of backstories corresponding to the same topic are used to further broaden the user query pool.

Crafty Crowdworkers. As noted in Section 2, it is well-known that a small but significant proportion of crowdworker responses are fraudulent. We were originally of the opinion that using an audio modality to solicit query variations would be resilient to such exploits. However the cleaning process identified submissions from machine-based *audio-to-text* models, and which could be confirmed by the lack of keystroke information. Some responses almost exactly matched the dictated topic, except with "speaker" annotations:

> Speaker 0: Two measures that sought to restrict fracking in Colorado won't appear on the ballot in November after ...

Similar cases were observed without the annotation, which may indicate that some workers assumed an audio transcription task, and replicated the summary. Although these particular cases were filtered out, the sophistication of the adversarial crowdworker responses was surprising, and provides a clear warning.

6 CONCLUSION AND FUTURE WORK

We have presented CC-News-En, a large-scale English news corpus derived from open-source Common Crawl data. While the original CC-News files can also be downloaded for free, reproducibly applying language filtering and pre-processing steps may be prohibitively expensive for some groups. Therefore, we provide the filtered collection as a series of compressed WARC files, further lowering the bar for reproducible experimentation. In addition to describing the methodology for generating the collection, we have collated a set of temporally matched crowdsourced queries.

[18] Private correspondence with Common Crawl Engineers.

In future work, we plan to remedy some of the shortcomings discussed in Section 5, including removal of non-news data, deduplication, and improved filtering of non-English documents. Perhaps the most useful extension to our work will be to gather relevance judgments, allowing the collection to be used as a fully-fledged Cranfield test collection. To achieve this goal, multiple diverse systems would need to be pooled across the queries associated with each of the topics, and followed by a round of relevance assessment. We hope that other research groups will assist in this task, to create an important and versatile shared resource that can serve the academic research community for a decade or more.

Resources. The corpus is available from http://go.unimelb.edu.au/u3nj. Tools for reproducing the corpus and related data are available from https://github.com/jmmackenzie/cc-news-tools/.

Acknowledgment. We thank Sebastian Nagel (Common Crawl) for providing useful information about the news crawl, and the Common Crawl organization for their commitment to providing open data. This work was partially supported by Australian Research Council Grants DP170102231, DP190101113, and DP200103136. The second author was supported by an RMIT VCPS.

REFERENCES

[1] A. Agarwal, A. Mandal, M. Schaffeld, F. Ji, J. Zhang, Y. Sun, and A. Aker. Good, neutral or bad: News classification. In *Proc. NewsIR'19 Workshop at SIGIR*, 2019.

[2] D. Albakour, M. Martinez, S. Tippmann, A. Aker, J. Stray, S. Dori-Hacohen, and A. Barrón-Cedeño. Third international workshop on recent trends in news information retrieval (NewsIR'19). In *Proc. SIGIR*, pages 1429–1431, 2019.

[3] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[4] J. Aslam, F. Diaz, M. Ekstrand-Abueg, R. McCreadie, V. Pavlu, and T. Sakai. TREC 2014 temporal summarization track overview. In *Proc. TREC*, 2014.

[5] L. Azzopardi and M. de Rijke. Automatic construction of known-item finding test beds. In *Proc. SIGIR*, pages 603–604, 2006.

[6] L. Azzopardi, R. W. White, P. Thomas, and N. Craswell. Data-driven evaluation metrics for heterogeneous search engine result pages. In *Proc. CHIIR*, pages 213–222, 2020.

[7] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User variability and IR system evaluation. In *Proc. SIGIR*, pages 625–634, 2015.

[8] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016.

[9] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*, pages 395–404, 2017.

[10] B. Baxendale. Machine-made index for technical literature – an experiment. *IBM Journal*, pages 354–361, 1958.

[11] R. Benham and J. S. Culpepper. Risk-reward trade-offs in rank fusion. In *Proc. Aust. Doc. Comp. Symp.*, pages 1.1–1.8, 2017.

[12] R. Benham, L. Gallagher, J. Mackenzie, T. T. Damessie, R.-C. Chen, F. Scholer, A. Moffat, and J. S. Culpepper. RMIT at the 2017 TREC CORE track. In *Proc. TREC*, 2017.

[13] R. Benham, L. Gallagher, J. Mackenzie, B. Liu, X. Lu, F. Scholer, A. Moffat, and J. S. Culpepper. RMIT at the 2018 TREC CORE track. In *Proc. TREC*, 2018.

[14] R. Benham, J. Mackenzie, A. Moffat, and J. S. Culpepper. Boosting search performance using query variations. *ACM Trans. Inf. Sys.*, 37(4):41.1–41.25, 2019.

[15] A. J. Biega, J. Schmidt, and R. S. Roy. Towards query logs for privacy studies: On deriving search queries from questions. In *Proc. ECIR*, pages 110–117, 2020.

[16] A. Checco, J. Bates, and G. Demartini. Adversarial attacks on crowdsourcing quality control. *J. Artif. Intell. Res.*, 67:375–408, 2020.

[17] M. Chmielewski and S. C. Kucker. An MTurk crisis? Shifts in data quality and the impact on study results. *Soc. Psychol. Pers. Sci.*, 11(4):464–473, 2020.

[18] A. Chuklin, A. Severyn, J. R. Trippas, E. Alfonseca, H. Silen, and D. Spina. Using audio transformations to improve comprehension in voice question answering. In *Proc. CLEF*, pages 164–170, 2019.

[19] R. Clancy, N. Ferro, C. Hauff, J. Lin, T. Sakai, and Z. Z. Wu. The SIGIR 2019 open-source IR replicability challenge (OSIRRC 2019). In *Proc. SIGIR*, pages 1432–1434, 2019.

[20] C. Cleverdon. The Cranfield tests on index language devices. *Aslib Proceedings*, 19(6):173–194, 1967.

[21] D. Corney, D. Albakour, M. Martinez, and S. Moussa. What do a million news articles look like? In *Proc. NewsIR'16 Workshop at ECIR*, pages 42–47, 2016.

[22] N. Ferro, N. Fuhr, M. Maistro, T. Sakai, and I. Soboroff. CENTRE@CLEF 2019. In *Proc. CLEF*, pages 283–290, 2019.

[23] J. S. Garofolo, E. M. Voorhees, C. G. Auzanne, V. M. Stanford, and B. A. Lund. 1998 TREC-7 spoken document retrieval track overview and results. In *Broadcast News Workshop*, pages 215–225, 1999.

[24] L. Han, K. Roitero, E. Maddalena, S. Mizzaro, and G. Demartini. On transforming relevance scales. In *Proc. CIKM*, pages 39–48, 2019.

[25] M. Hu, A. Sun, and E.-P. Lim. Comments-oriented blog summarization by sentence extraction. In *Proc. CIKM*, pages 901–904, 2007.

[26] U. Krishnan, B. Billerbeck, A. Moffat, and J. Zobel. Abstraction of query auto completion logs for anonymity-preserving analysis. *Inf. Retr.*, 22(5):499–524, 2019.

[27] U. Krishnan, B. Billerbeck, A. Moffat, and J. Zobel. Generation of synthetic query auto completion logs. In *Proc. ECIR*, pages 621–635, 2020.

[28] C.-Y. Lin and E. Hovy. Identifying topics by position. In *Proc. ANLP*, pages 283–290, 1997.

[29] J. Lin, M. Crane, A. Trotman, J. Callan, I. Chattopadhyaya, J. Foley, G. Ingersoll, C. Macdonald, and S. Vigna. Toward reproducible baselines: The open-source IR reproducibility challenge. In *Proc. ECIR*, 2016.

[30] J. Lin, A. Roegiest, L. Tan, R. McCreadie, E. Voorhees, and F. Diaz. Overview of the TREC 2016 real-time summarization track. In *Proc. TREC*, 2016.

[31] J. Lin, J. Mackenzie, C. Kamphuis, C. Macdonald, A. Mallia, M. Siedlaczek, A. Trotman, and A. de Vries. Supporting interoperability between open-source search engines with the common index file format. In *Proc. SIGIR*, pages 2149–2152, 2020.

[32] B. Liu, N. Craswell, X. Lu, O. Kurland, and J. S. Culpepper. A comparative analysis of human and automatic query variants. In *Proc. ICTIR*, pages 47–50, 2019.

[33] J. Mackenzie, K. Gupta, F. Qiao, A. H. Awadallah, and M. Shokouhi. Exploring user behavior in email re-finding tasks. In *Proc. WWW*, pages 1245–1255, 2019.

[34] J. Mackenzie, A. Mallia, M. Petri, J. S. Culpepper, and T. Suel. Compressing inverted indexes with recursive graph bisection: A reproducibility study. In *Proc. ECIR*, pages 339–352, 2019.

[35] S. Mackie, R. McCreadie, C. Macdonald, and I. Ounis. Experiments in newswire summarisation. In *Proc. ECIR*, pages 421–435, 2016.

[36] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proc. SIGIR*, pages 135–144, 2017.

[37] A. Moffat. Judgment pool effects caused by query variations. In *Proc. Aust. Doc. Comp. Symp.*, pages 65–68, 2016.

[38] A. Moffat and M. Petri. Index compression using byte-aligned ANS coding and two-dimensional contexts. In *Proc. WSDM*, pages 405–413, 2018.

[39] F. Moraes, J. Yang, R. Zhang, and V. Murdock. The role of attributes in product quality comparisons. In *Proc. CHIIR*, pages 253–262, 2020.

[40] A. Nenkova. Automatic text summarization of newswire: Lessons learned from the Document Understanding Conference. In *Proc. AAAI*, pages 1436–1441, 2005.

[41] M. Petri and A. Moffat. Compact inverted index storage using general-purpose compression libraries. *Soft. Prac. & Exp.*, 48(4):974–982, 2018.

[42] G. E. Pibiri and R. Venturini. On optimally partitioning variable-byte codes. *IEEE Trans. Knowl. Data Eng.*, 2019.

[43] G. E. Pibiri, M. Petri, and A. Moffat. Fast dictionary-based compression for inverted indexes. In *Proc. WSDM*, pages 6–14, 2019.

[44] M. Sanderson. Test collection based evaluation of information retrieval systems. *Found. Trnd. Inf. Retr.*, 4(4):247–375, 2010.

[45] I. Soboroff, S. Huang, and D. Harman. TREC 2018 news track overview. In *Proc. TREC*, 2018.

[46] D. Spina, J. R. Trippas, L. Cavedon, and M. Sanderson. Extracting audio summaries to support effective spoken document search. *J. Assoc. Inf. Sci. Technol.*, 68(9): 2101–2115, 2017.

[47] I. Stanton, S. Ieong, and N. Mishra. Circumlocution in diagnostic medical queries. In *Proc. SIGIR*, pages 133–142, 2014.

[48] J. R. Trippas, D. Spina, M. Sanderson, and L. Cavedon. Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proc. SIGIR*, pages 991–994, 2015.

[49] P. Yang, H. Fang, and J. Lin. Anserini: Reproducible ranking baselines using Lucene. *J. Data Inf. Qual.*, 10(4):1–20, 2018.

[50] J. Ye and S. Skiena. Mediarank: Computational ranking of online news sources. In *Proc. KDD*, pages 2469–2477, 2019.