A Massive Scale Semantic Similarity Dataset of Historical English

Emily Silcock[1], **Melissa Dell**[1,2*]
[1]Harvard University; Cambridge, MA, USA.
[2]National Bureau of Economic Research; Cambridge, MA, USA.
[*]Corresponding author: melissadell@fas.harvard.edu.

Abstract

A diversity of tasks use language models trained on semantic similarity data. While there are a variety of datasets that capture semantic similarity, they are either constructed from modern web data or are relatively small datasets created in the past decade by human annotators. This study utilizes a novel source, newly digitized articles from off-copyright, local U.S. newspapers, to assemble a massive-scale semantic similarity dataset spanning 70 years from 1920 to 1989 and containing nearly 400M positive semantic similarity pairs. Historically, around half of articles in U.S. local newspapers came from newswires like the Associated Press. While local papers reproduced articles from the newswire, they wrote their own headlines, which form abstractive summaries of the associated articles. We associate articles and their headlines by exploiting document layouts and language understanding. We then use deep neural methods to detect which articles are from the same underlying source, in the presence of substantial noise and abridgement. The headlines of reproduced articles form positive semantic similarity pairs. The resulting publicly available `HEADLINES` dataset is significantly larger than most existing semantic similarity datasets and covers a much longer span of time. It will facilitate the application of contrastively trained semantic similarity models to a variety of tasks, including the study of semantic change across space and time.

1 Introduction

Transformer language models contrastively trained on large-scale semantic similarity datasets are integral to a variety of applications in natural language processing (NLP). Contrastive training is often motivated by the anisotropic geometry of pre-trained transformer language models like BERT and GPT [7], which complicates working with their hidden representations. Representations of low frequency words are pushed outwards, the sparsity of low frequency words violates convexity, and the distance between embeddings is correlated with lexical similarity. This leads to poor alignment between semantically similar texts and poor performance when individual term representations from the transformer model are pooled to create a representation for longer texts, since pooling assumes convexity. Contrastive training reduces anisotropy, increasing uniformity of the distribution of features on the hypersphere and improving alignment of features from positive pairs [26]. This in turn significantly improves pooled representations of sentences (or longer documents) [23].

A variety of semantic similarity datasets have been used for contrastive training. Many of these datasets are relatively small, and the bulk of the larger datasets are created from recent web texts; *e.g.*, positive pairs are drawn from the texts in an online comment thread or from questions marked as duplicates in a forum. To fill important gaps in the suite of existing datasets, this study develops `HEADLINES` (**H**istorical **E**normous-Scale **A**bstractive **D**up**LI**cate **Ne**ws **S**ummaries), a massive dataset containing nearly 400M high quality semantic similarity pairs drawn from 70 years of off copyright

Preprint. Under review.

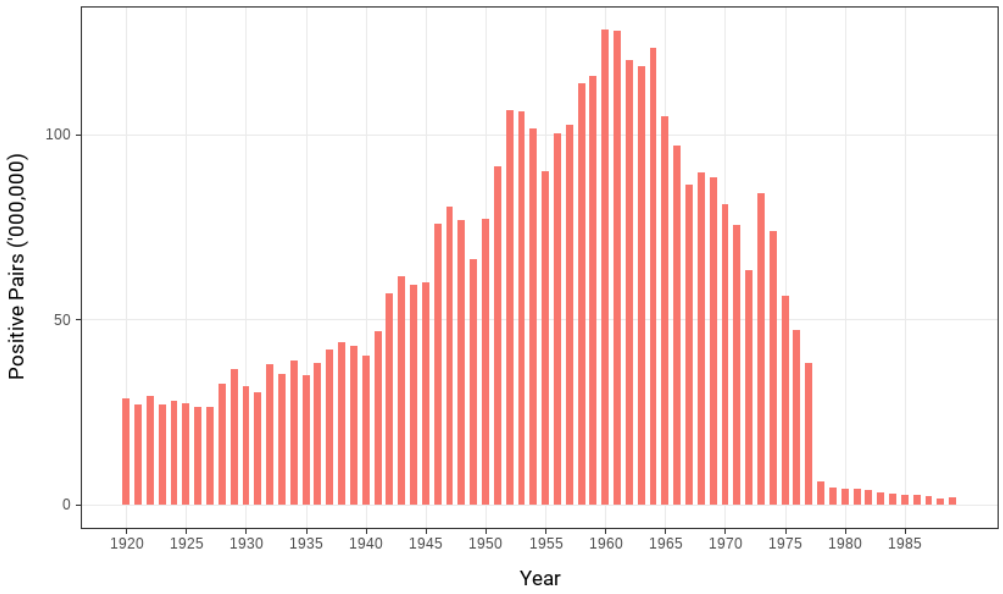


Figure 1: Semantic similarity pairs by year.

U.S. newspapers. Historically, around half of content in the many thousands of local newspapers across the U.S. was taken from centralized sources such as the Associated Press [11]. Local newspapers reprinted wire articles but wrote their own headlines, which form abstractive summaries of the articles. Headlines written by different papers to describe the same wire article form positive semantic similarity pairs.

To construct HEADLINES, we digitize front pages of off-copyright newspaper page scans, localizing and OCRing individual content regions like headlines and articles. The headlines, bylines, and article texts that form full articles span multiple bounding boxes - often arranged with complex layouts - and we associate them using a model that combines layout information and language understanding [20]. Then, we use neural methods from [24] to accurately predict which articles come from the same underlying source, in the presence of noise and abridgement.

HEADLINES allows us to leverage the collective writings of many thousands of local editors across the U.S., spanning much of the 20th century, to create a massive, high-quality dataset that captures semantic similarity across time. The dataset contains 393,635,650 positive semantic similarity pairs, spanning the period between 1920 and 1989 and covering 49 of the 50 states (Figure 1). The dataset could be significantly expanded by extending the pipeline to interior newspaper pages, which also published extensive syndicated content. We do not do so at this time, as the dataset is already massive.

Relative to existing datasets, HEADLINES is very large, covering a vast array of topics. This makes it useful generally speaking for semantic similarity pre-training. It also covers a long period of time, making it a rich training data source for the study of historical texts and semantic change. It captures semantic similarity directly, as the positive pairs summarize the same underlying texts.

This study is organized as follows. Section 2 describes HEADLINES, and Section 3 discusses how it relates to existing datasets. Section 4 outlines the methods used for dataset construction, and Section 5 concludes.

2 Dataset Description

HEADLINES contains 393,635,650 positive headline pairs from off-copyright newspapers. Figure 2 shows examples of semantic similarity pairs, which include noise from the OCR. We do not post-process the headlines with spellchecking, to provide users maximal flexibility in choosing how aggressive they wish to be with spellchecking.

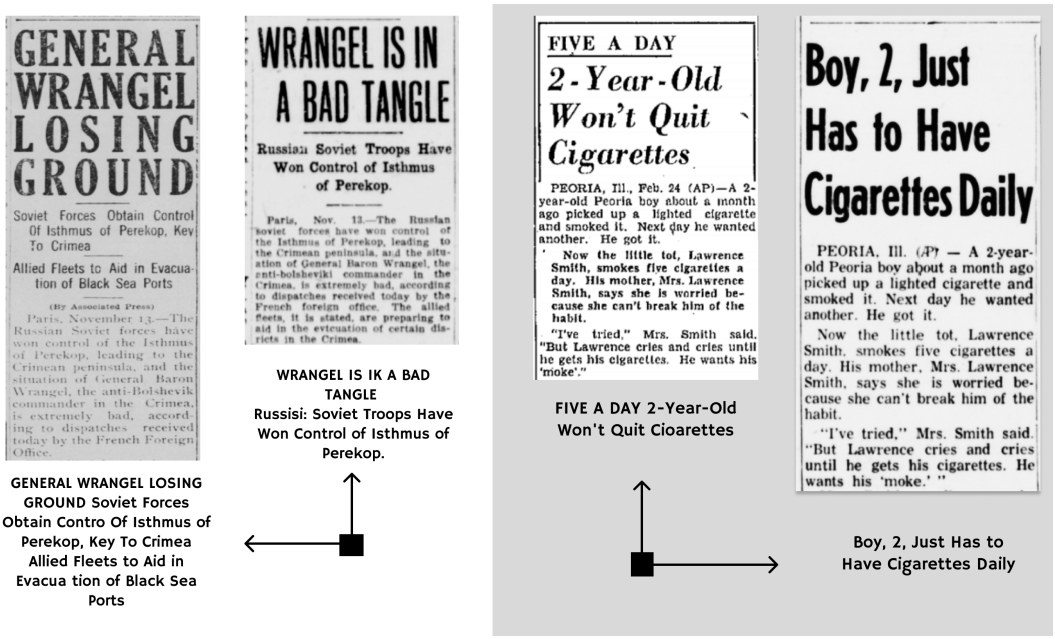


Figure 2: Semantic similarity examples, showing article image crops and OCR'ed headlines.

Figure 1 plots the number of headline pairs by year, illustrating coverage across time from 1920 to 1989. Content declines sharply in the late 1970s, due to a major copyright law change effective January 1, 1978.

Figure 3 plots the distribution of content by state, illustrating national coverage. Newspapers covered a vast array of topics, leading to strong coverage across the vocabulary.

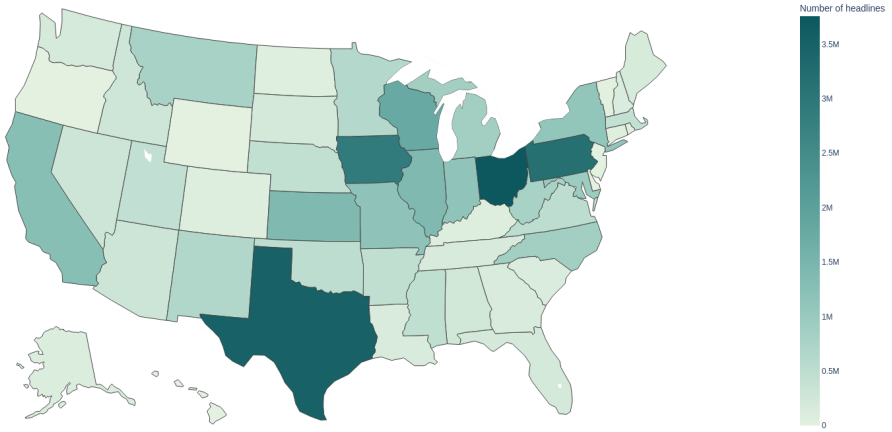


Figure 3: Geographic variation in source of headlines

Content are drawn from off-copyright newspapers. Earlier in the 20th century, it was rare for papers outside those with the widest circulation to include a copyright notice (or to renew copyrights if published with notice), required formalities at the time. Individual pieces of syndicated content in off-copyright papers could be copyrighted: most frequently comics, occasionally ads - these are detected by our layout model and we do not OCR either - and occasional runs of syndicated fiction. News wire articles were not copyrighted [21], as renewing a copyright on yesterday's news would

have been costly with no commercial value, and in any case our dataset exploits locally written headlines. Headlines are a small share of overall text content.

HEADLINES has a Creative Commons CC-BY license, to encourage widespread use, and is available on Huggingface.[1]

3 Existing Semantic Similarity Datasets

There is a dense literature on semantic similarity, with datasets covering diverse types of textual similarity and varying greatly in size. The focus of HEADLINES on semantic similarity in historical texts sets it apart from other widely used datasets. It also dwarfs the size of most existing datasets, aggregating the collective work of 20th century newspapers editors, from towns across the U.S. Its paired headlines summarize the same text, rather than being related by other forms of similarity frequently captured by datasets, such as being in the same conversation thread or answering a corresopnding question.

Amongst the most closely related semantic similarity datasets to HEADLINES are Microsoft COCO [4] and Flickr 30k [28].[2] MS COCO used Amazon's Mechanical Turk to collect five captions for each image in the dataset, resulting in around 828,000 positive caption pairs. In Flickr, around 152,000 crowd-sourced captions describing around 32,000 images yield 317,695 positive semantic similarity pairs. Like HEADLINES, positive pairs in these datasets describe the same underlying content, but are describing an image rather than providing an abstractive summary of a longer text. The nearly 400M semantic similarity pairs in HEADLINES significantly exceed the size of these datasets, which are used for a much broader variety of tasks including object detection and semantic segmentation. In future work, HEADLINES could be expanded to include caption pairs describing the same underlying photo wire image, as local papers frequently wrote their own captions. We have developed a pipeline for detecting reproduced images [1] that would yield hundreds of millions of positive semantic similarity pairs describing images rather than texts. We do not deploy it for this iteration of HEADLINES, since the dataset already contains nearly 400M positive pairs and processing this content is costly.

Another related class of semantic similarity datasets consists of duplicate questions from web platforms. These datasets include questions tagged by users as duplicates from WikiAnswers (77.4 million positive pairs) [8], duplicate stack exchange questions (around 304,000 duplicate title pairs) [2], and duplicate quora questions (around 400,000 duplicate pairs) [15]. These datasets are created from recent web texts, and hence cannot be used to study semantic change across long spans of time. They also do not contain systematic information about the geographic location of the author.

Online comment threads have also been used to train semantic similarity models. For example, the massive scale Reddit Comments [14] draws positive semantic similarity pairs from Reddit conversation threads between 2016 and 2018, providing 726.5 million positive pairs. Semantic similarity between comments in an online thread reflects conversational similarity, to the extent the thread stays on topic, rather than abstractive similarity, making HEADLINES a complement.

While other datasets exploit abstractive summaries, to our knowledge there are not large-scale datasets with abstractive summaries of the same underlying texts. Semantic Scholar (S2ORC), a database of academic articles, has been used to create semantic similarity pairs of the titles and abstracts of papers that cite each other. For instance, Sentence BERT [23] trains on 52.6 million positive citation abstract pairs constructed from S2ORC, capturing topic similarity.

Finally, question-answer and natural-language inference datasets are widely used for semantic similarity training. These include 65 million question-answer pairs from PAQ [19], 9 million MS MARCO question-answer triples [5], 3 million question answer pairs for GOOAQ [17], 1.2 million title-answer pairs from Yahoo Answers [29], 582,000 question-answer pairs from SearchQA [6], 100,000 question-answer pairs from Natural Questions [18], 87,000 question-answer pairs from SQuAD [22], and 277,000 NLI pairs from SNLI [3] and MultiNLI [27]. These datasets capture the semantic similarity between questions and answers or hypotheses and premises - rather than abstractive similarity - again making them complements to HEADLINES.

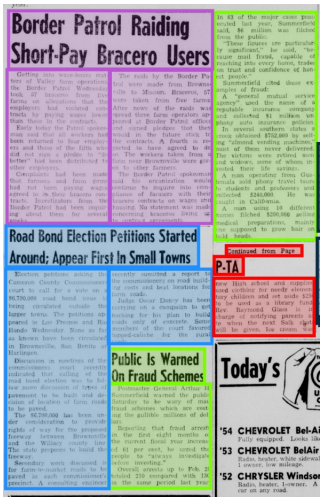
[1] https://huggingface.co/datasets/dell-research-harvard/headlines-semantic-similarity

[2] Dataset sizes are drawn, when applicable, from a table documenting the training of Sentence BERT [23].

4 Dataset Construction

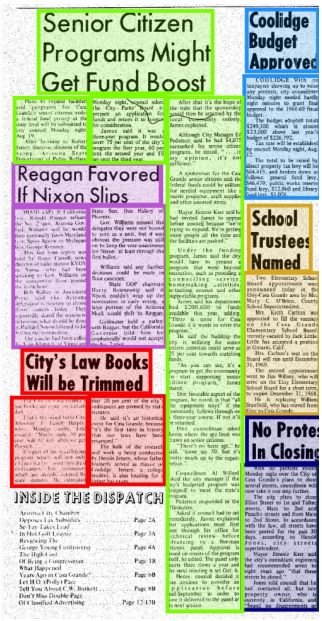
We digitized front pages from off-copyright newspapers spanning 1920-1989. We recognize layouts using Mask RCNN [13] and use Tesseract to OCR the texts.

After digitization, we associate multiple bounding boxes comprising the same article. We associate the (potentially multiple) headline bounding boxes with the (potentially multiple) article bounding boxes and byline boxes using a combination of layout information and language understanding.



(a) The green article is non-contiguous and it is unclear if the second part continues the purple, green, or even blue articles if the text is not used.

(b) The green article is nested inside the blue article, which causes rule-based and image-based models to either associate the green and blue articles together, or miss the second half of the blue article.



(c) The smallest rectangle that contains the green article also contains the purple and red articles.

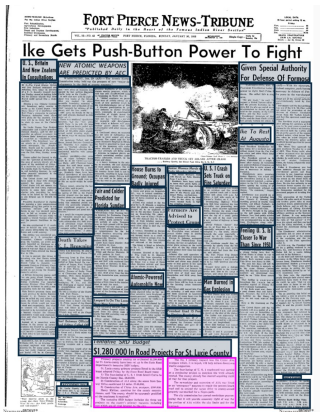
Figure 4: Articles that are misassociated with rule-based or image-based methods

A rule-based approach using the document layouts gets many of the associations correct, but misses some difficult cases where article bounding boxes are arranged in complex layouts, with Figure 4 providing representative examples. Language understanding can be used to associate such articles but must be robust to noise, from errors in layout detection (*e.g.* from cropping part of a content bounding box or adding part of the line below), from errors in Tesseract's rule-based line detection algorithm, and from OCR character recognition errors.

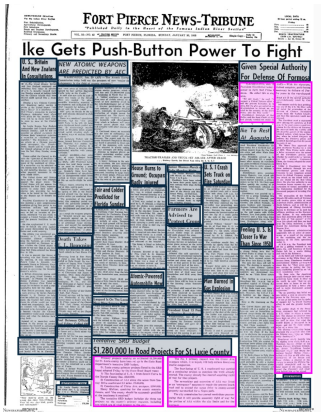
Hand annotating a sufficiently large training dataset would have been too costly. Instead, we devise a set of rules that - while recall is relatively low - have nearly perfect precision, as measured on an evaluation set. The algorithm exploits the positioning of article bounding boxes relative to headline boxes (as in Figure 5), first grouping an article bounding box with a headline bounding box if the rules are met and then associating all article bounding boxes grouped with the same headline together. Since precision is above 0.99, the rule can be used to generate silver-quality training data.

To train a classifier on these data to predict whether article box B follows box A, we embed the first and last 64 tokens of the texts in boxes B and A, respectively, with a RoBERTa base model [20]. The pair is positive when B follows A. The training set includes 12,769 positive associated pairs, with training details described in the supplementary materials.

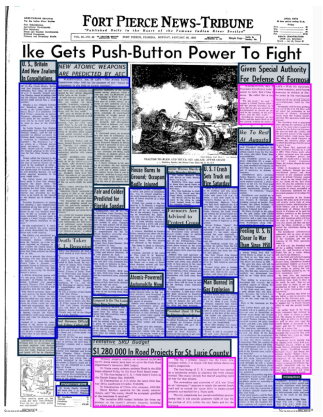
At inference time, we first associate texts using the rule-based approach, described in Figure 5, which has extremely high precision. To improve recall, we then apply the RoBERTa cross-encoder to remaining article boxes that could plausibly be associated, given their coordinates. Texts cannot be followed by a text that appears to the left by a sufficiently wide margin, as layouts - while sometimes complex - proceed from left to right. Hence, these combinations are not considered.



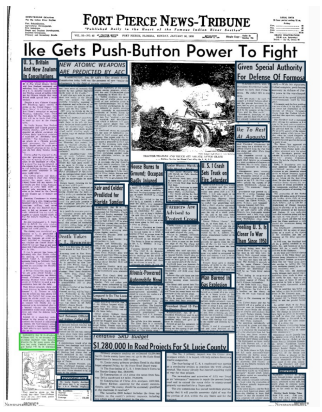
(a) Article bounding boxes that are under the same headline and at the bottom of the page are used to create training data.



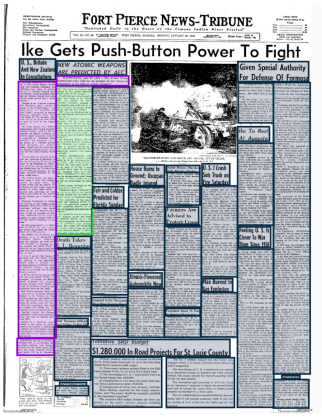
(b) At inference time, all bounding boxes that are directly under the same headline are associated with each other.



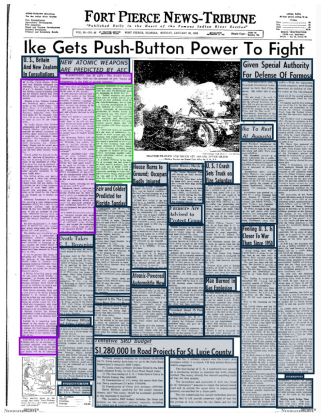
(c) Any other article with a headline directly above it is not compared, leaving only a few orphans that are left to be associated.



(d) This orphan (green bounding box) is compared with the bounding box above it. In this case, it is a separate article without a headline so it is not associated.



(e) Orphans are also compared with bounding boxes to the right. This orphan (green bounding box) is associated with the bounding box directly above.



(f) The final orphan is compared with two columns to the right, as sometimes articles skip columns. If further columns to the right exist, they are not compared.

Figure 5: Illustration of article association pipeline

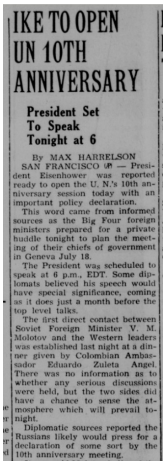
Table 1 shows recall, precision and F1 for associated articles when the neural model is used in combination with the rule based approach. The F1 of 93.7 is high, and precision is extremely high. Errors typically occur when there is an error in the layout analysis model or when contents are very similar, for instance, grouping multiple obituaries into a single article.

	(1) F1	(2) Recall	(3) Precision
Full Article Association	93.7	88.3	99.7

Table 1: This table evaluates the full article association model, showing F1, recall, and precision, respectively.

Accurately detecting reproduced content can be challenging, as articles were often heavily abridged by local papers to fit within their space constraints and errors in OCR or article association can add significant noise. Table 2 shows examples of reproduced articles.

We use the model developed by [24], who show that a contrastively trained neural MPNet bi-encoder - combined with single linkage clustering of article representations - accurately and cheaply detects


IKE TO OPEN UN 10TH ANNIVERSARY

President Set To Speak Tonight at 6

By MAX HARRELSON

SAN FRANCISCO (AP — President Eisenhower was reported ready to open the U. N.'s 10th anniversary session today with an important policy declaration.

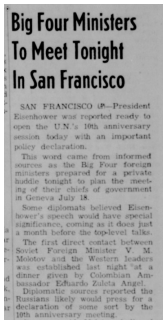
This word came from informed sources as the Big Four foreign ministers prepared for a private huddle tonight to plan the meeting of their chiefs of government in Geneva July 18.

The President was scheduled to speak at 6 p.m., EDT. Some diplomats believed his speech would have special significance, coming as it does just a month before the top level talks.

The first direct contact between Soviet Foreign Minister V. M. Molotov and the Western leaders was established last night at a dinner given by Colombian Ambassador Eduardo Zuleta Angel. There was no information as to whether any serious discussions were held, but the two sides did have a chance to sense the atmosphere which will prevail tonight.

Diplomatic sources reported the Russians likely would press for a declaration of some sort by the 10th anniversary meeting.

SAN FRANCISCO (# — Presi-dent Eisenhower was_ reported ready to open the U. N.'s 10th anniversary session today with an important policy declaration.
This word came from informed sources as the Big Four foreign ministers prepared for a private huddle tonight to plan the meeting of their chiefs of government in Geneva July 18.
The President was scheduled to speak at 6 p.m., EDT. Some diplomats believed his speech would have special significance, coming as it does just a month before the top level talks.
The first direct contact between Soviet Foreign Minister V. M. Molotov and the Western leaders was established last night at a dinner given by Colombian Ambassador Eduardo Zuleta Angel There was no information as tc whether any serious discussion: were held, but the two sides dic have a chance to sense the at-'}mosphere which will prevail to-'l night..l
Diplomatic sources reported th« Russians likely would press for : 'Ideclaration of some sort by the | 10th anniversary meeting


Big Four Ministers To Meet Tonight In San Francisco

SAN FRANCISCO (AP—President Eisenhower was reported ready to open the U.N.'s 10th anniversary session today with an important policy declaration.

This word came from informed sources as the Big Four foreign ministers prepared for a private huddle tonight to plan the meeting of their chiefs of government in Geneva July 18.

Some diplomats believed Eisenhower's speech would have special significance, coming as it does just a month before the top-level talks.

The first direct contact between Soviet Foreign Minister V. M. Molotov and the Western leaders was established last night "at a dinner given by Colombian Ambassador Eduardo Zuleta Angel.

Diplomatic sources reported the Russians likely would press for a declaration of some sort by the 10th anniversary meeting.

SAN FRANCISCO #—President Eisenhower was reported ready to open the U_N.'s 10th anniversary session today with an important policy declaration.
This word came from informed sources as the Big Four foreign ministers prepared for a private huddle tonight to plan the meeting of their chiefs of government in Geneva July 18.
Some diplomats believed Eisenhower's speech would have special significance, coming as it does just 'a month before the top-level talks.| The first direct contact between (Soviet Foreign Minister V. M.| Molotov and the Western leaders ¦ was established last night "at a'dinner given by Colombian Aml bassador Eduardo Zulcta Angel.
¦ Diplomatic sources reported the 'Russians likely would press for a declaration of some sort by the 110th anniversary meeting.


Ike's Talk Will Open U. N. Meet

Big 4 Ministers Arrange Private Huddle Tonight

By MAX HARRELSON

SAN FRANCISCO (AP—President Eisenhower was reported ready to open the U.N.'s 10th anniversary session today with an important policy declaration.

This word came from informed sources as the Big Four foreign ministers prepared for a private huddle tonight to plan the meeting of their chiefs of government in Geneva July 18.

Some diplomats believed Eisenhower's speech would have special significance, coming as it does just a month before the top-level talks.

Mulles to Bo Host

The first direct contact between Soviet Foreign Minister V. M. Molotov and the Western leaders was established last night at a dinner given by Colombian Ambassador Eduardo Zuleta Angel.

Diplomatic sources reported the Russians likely would press for a declaration of some sort by the 10th anniversary meeting.

"SY MAA MARKRKRELSUN
SAN FRANCISCO #®—President Eisenhower was reported ready to open the U.N.'s 10th anniversary session today with an importan! policy declaration.
This word came from informed sources as the Big Four foreign ministers prepared for a private huddle tonight to plan the meeting of their chiefs of government in Geneva July 18.
Some diplomats believed Eisen.hower's speech would have special! significance, coming as it does just a month before the top-level talks.
Mulles to Bo Host
The first direct contact between Soviet Foreign Minister V. M. Molotov and the Western leaders was established last night at a dinner given by Colombian Ambassador Eduardo Zuleta Angel.
Diplomatic sources reported the Russians likely would press for a 'declaration of some sort by the 10th anniversary meeting.

Table 2: Examples of reproduced articles. Additions are highlighted, and OCR errors are underlined.

reproduced content. This bi-encoder is contrastively trained on a hand-labeled dataset to create similar representations of articles from the same wire source and dissimilar representations of articles from different underlying sources, using S-BERT's online contrastive loss [12] implementation.

On a labeled sample of all front page articles appearing in our newspaper corpus for three days in the 1930s and 1970s, the neural bi-encoder methods achieve an adjusted rand index of 91.5, compared to 73.7 for an optimal local sensitive hashing specification chosen on the validation set (another full-day labeled sample) [24]. The pipeline can be scaled to 10 million articles on a single GPU in a matter of hours, using a FAISS [16] backend. Errors typically consist of articles about the same story from different news wire services (*e.g.* the Associated Press and the United Press) or updates to a story as new events unfolded. Both types of errors will plausibly still lead to informative semantic similarity pairs.

We run clustering on the article embeddings by year over all years in our sample. In post-processing, we use a simple set of rules exploiting the dates of articles within clusters to remove content like weather forecasts and legal notices, that are highly formulaic and sometimes cluster together when they contain very similar content (*e.g.* a similar 5 day forecast) but did not actually come from the same underlying source. We remove all headline pairs that are below a Levenshtein edit distance, normalized by the min length in the pair, of 0.1, to remove pairs that are exact duplicates up to OCR noise. More details are provided in the supplementary materials.

The digitization of newspaper scans was performed using Azure F-Series CPU nodes. Full article association and detection of reproduced content were performed on an A6000 GPU card. All training was conducted on an A6000 GPU card.

5 Conclusion

HEADLINES provides a massive-scale semantic similarity dataset, consisting of nearly 400M positive semantic similarity pairs. It combines the collective work of thousands of local newspaper editors across much of the 20th century, providing rich data that can be useful more generally for contrastively training large language models and specifically for studying semantic change. HEADLINES forms part of a broader agenda to create more resources for deploying large language models on historical texts.

Supplementary Materials

S-1 Methods to Associate Articles

To detect whether article bounding box B follows box A, we train a cross-encoder using a RoBERTa base model [20]. Figure 5, Panel A in the main text describes how positive pairs are created. For hard negatives, we used article boxes under the same headline in reverse reading order (right to left). For standard negatives, we took pairs of articles on the same page, where B was above and to the left of A, as articles do not read from right to left. One twelfth of our training data were positive pairs, another twelfth were hard negative pairs and the remainder were standard negative pairs. This outperformed a more balanced training sample.

We used a Bayesian search algorithm [9] to find optimal hyperparameters on one tenth of our training data (limited compute prevented us from running this search with the full dataset), which led to a learning rate of 1.7e-5, with a batch size of 64 and 29.57% warm up. We trained for 26 epochs with an AdamW optimizer, and optimize a binary cross-entropy loss.

S-2 Methods to Detect Reproduced Content

To detect reproduced content, we use the contrastively trained bi-encoder model developed by [24], which is trained to learn similar representations for reproduced articles and dissimilar representations for non-reproduced articles. This model is based on an S-BERT MPNET model [23, 25] and is fine-tuned on a hand-labelled dataset of articles from the same underlying wire source, using S-BERT's online contrastive loss [12] implementation, with a 0.2 margin and cosine similarity as the distance metric. The learning rate is 2e-5 with 100% warm up and a batch size of 32. It uses an AdamW optimizer, and the model is trained for 16 epochs.

To create clusters from the bi-encoder embeddings, we use highly scalable single-linkage clustering, with a cosine similarity threshold of 0.94. We build a graph using articles as nodes, and add edges if the cosine similarity is above this threshold. As edge weights we use the negative exponential of the difference in dates (in days) between the two articles. We then apply Leiden community detection to the graph to control false positive edges that can otherwise merge disparate groups of articles.

We further remove clusters that have over 50 articles and contain articles with greater than five different dates. We also remove clusters that contain over 50 articles, when the number of articles is more than double the number of unique newspapers from which these articles are sourced. This removes clusters of content that are correctly clustered in the sense of being based on the same underlying source, but are not useful for the `HEADLINES` dataset. For example, an advertisement (misclassified as an article due to an article-like appearance) might be repeated by the same newspaper on multiple different dates and would be removed by these rules, or weather forecasts can be very near duplicates across space and time, forming large clusters.

S-3 Dataset Information

S-3.1 Dataset URL

`HEADLINES` can be found at `https://huggingface.co/datasets/dell-research-harvard/headlines-semantic-similarity`.

This dataset has structured metadata following `schema.org`, and is readily discoverable.[3]

S-3.2 DOI

The DOI for this dataset is: 10.57967/hf/0751.

[3] See `https://search.google.com/test/rich-results/result?id=_HKjxIv-LaF_8ElAarsM_g` for full metadata.

S-3.3 License

HEADLINES has a Creative Commons CC-BY license.

S-3.4 Dataset usage

The dataset is hosted on huggingface, in json format. Each year in the dataset is divided into a distinct file (eg. 1952_headlines.json).

The data is presented in the form of clusters, rather than pairs to eliminate duplication of text data and minimize the storage size of the datasets.

An example from HEADLINES looks like:

```
{
    "headline": "FRENCH AND BRITISH BATTLESHIPS IN MEXICAN WATERS",
    "group_id": 4
    "date": "May-14-1920",
    "state": "kansas",
}
```

The data fields are:

- `headline`: headline text.
- `date`: the date of publication of the newspaper article, as a string in the form mmm-DD-YYYY.
- `state`: state of the newspaper that published the headline.
- `group_id`: a number that is shared with all other headlines for the same article. This number is unique across all year files.

The whole dataset can be easily downloaded using the `datasets` library:

```
from datasets import load_dataset
dataset_dict = load_dataset("dell-research-harvard/headlines-semantic-
similarity")
```

Specific files can be downloaded by specifying them:

```
from datasets import load_dataset
load_dataset(
    "dell-research-harvard/headlines-semantic-similarity",
    data_files=["1929_headlines.json", "1989_headlines.json"]
)
```

S-3.5 Author statement

We bear all responsibility in case of violation of rights.

S-3.6 Maintenance Plan

We have chosen to host HEADLINES on huggingface as this ensures long-term access and preservation of the dataset.

S-3.7 Dataset documentation and intended uses

We follow the datasheets for datasets template [10].

S-3.7.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Transformer language models contrastively trained on large-scale semantic similarity datasets are integral to a variety of applications in natural language processing (NLP). A variety of semantic similarity datasets have been used for this purpose, with positive text pairs related to each other in some way. Many of these datasets are relatively small, and the bulk of the larger datasets are created from recent web texts; e.g. positives are drawn from the texts in an online comment thread or duplicate questions in a forum. Relative to existing datasets, `HEADLINES` *is very large, covering a vast array of topics. This makes it useful generally speaking for semantic similarity pre-training. It also covers a long period of time, making it a rich training data source for the study of historical texts and semantic change. It captures semantic similarity directly, as the positive pairs summarize the same underlying texts.*

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
`HEADLINES` *was created by Melissa Dell and Emily Silcock, at Harvard University.*

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of the dataset was funded by the Harvard Data Science Initiative, Harvard Catalyst, and compute credits provided by Microsoft Azure to the Harvard Data Science Initiative.

Any other comments?
None.

S-3.7.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

`HEADLINES` *comprises instances of newspaper headlines and relationships between them. Specifically, each headline includes information on the text of the headline, the date of publication, and the state it was published in. Headlines have relationships between them if they are semantic similarity pairs, that is, if they two different headlines for the same newspaper article.*

How many instances are there in total (of each type, if appropriate)?
`HEADLINES` *contains 34,867,488 different headlines and 396,001,930 positive relationships between headlines.*

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Many local newspapers were not preserved, and newspapers with the widest circulation tended to renew their copyrights, so cannot be included.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each data instance consists of raw data. Specifically, an example from `HEADLINES` *is:*


```
{
    "headline": "FRENCH AND BRITISH BATTLESHIPS IN MEXICAN WATERS",
```

```
    "group_id": 4
    "date": "May-14-1920",
    "state": "kansas",
}
```

The data fields are:

- `headline`*: headline text.*
- `date`*: the date of publication of the newspaper article, as a string in the form mmm-DD-YYYY.*
- `state`*: state of the newspaper that published the headline.*
- `group_id`*: a number that is shared with all other headlines for the same article. This number is unique across all year files.*

Is there a label or target associated with each instance? If so, please provide a description.

Each instance contains a `group_id` *as mentioned directly above. This is a number that is shared by all other instances that are positive semantic similarity pairs.*

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

In some cases, the state of publication is missing, due to incomplete metadata.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between instances are made explicit in the `group_id` *variable, as detailed above.*

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no recommended splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The data is sourced from OCR'd text of historical newspapers. Therefore some of the headline texts contain OCR errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The dataset does not contain information that might be viewed as confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

The headlines in the dataset reflect diverse attitudes and values from the period in which they were written, 1920-1989, and contain content that may be considered offensive for a variety of reasons.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Many news articles are about people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset does not specifically identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

If an individual appeared in the news during this period, then headline text may contain their name, age, and information about their actions.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

All information that it contains is already publicly available in the newspapers used to create the headline pairs.

Any other comments?
None.

S-3.7.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

To create HEADLINES, we digitized front pages from off-copyright newspapers spanning 1920-1989. Historically, around half of articles in U.S. local newspapers came from newswires like the Associated Press. While local papers reproduced articles from the newswire, they wrote their own headlines, which form abstractive summaries of the associated articles. We associate articles and their headlines by exploiting document layouts and language understanding. We then use deep neural methods to detect which articles are from the same underlying source, in the presence of substantial noise and abridgement. The headlines of reproduced articles form positive semantic similarity pairs.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

These methods are described in detail in the main text and supplementary materials of this paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The dataset was not sampled from a larger set.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
We used student annotators to create the validation sets for associating bounding boxes, and the training and validation sets for clustering duplicated articles. They were paid $15 per hour, a rate set by a Harvard economics department program providing research assistantships for undergraduates.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The headlines were written between 1920 and 1989. Semantic similarity pairs were computed in 2023.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, this dataset uses entirely public information and hence does not fall under the domain of Harvard's institutional review board.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Historical newspapers contain a variety of information about people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The data were obtained from off-copyright historical newspapers.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Individuals were not notified; the data came from publicly available newspapers.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The dataset was created from publicly available historical newspapers.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?
None.

S-3.7.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

See the description in the main text.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No specific software was used to clean the instances.

Any other comments?
None.

S-3.7.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?
The dataset can be used for training models for semantic similarity, studying language change over time and studying difference in language across space.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The dataset contains historical news headlines, which will reflect current affairs and events of the time period in which they were created, 1920-1989, as well as the biases of this period.

Are there tasks for which the dataset should not be used? If so, please provide a description.

It is intended for training semantic similarity models and studying semantic variation across space and time.

Any other comments?
None

S-3.7.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is available for public use.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is hosted on huggingface. Its DOI is 10.57967/hf/0751.

When will the dataset be distributed?
The dataset was distributed on 7th June 2023.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is distributed under a Creative Commons CC-BY license. The terms of this license can be viewed at `https://creativecommons.org/licenses/by/2.0/`

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

There are no third party IP-based or other restrictions on the data.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No export controls or other regulatory restrictions apply to the dataset or to individual instances.

Any other comments?
None.

S-3.7.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted on huggingface.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The recommended method of contact is using the huggingface 'community' capacity. Additionally, Melissa Dell can be contacted at melissadell@fas.harvard.edu.

Is there an erratum? If so, please provide a link or other access point.

There is no erratum.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

We have no plans to update the dataset. If we do, we will notify users via the huggingface Dataset Card.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

There are no applicable limits on the retention of data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

We have no plans to update the dataset. If we do, older versions of the dataset will not continue to be hosted. We will notify users via the huggingface Dataset Card.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Others can contribute to the dataset using the huggingface 'community' capacity. This allows for anyone to ask questions, make comments and submit pull requests. We will validate these pull requests. A record of public contributions will be maintained on huggingface, allowing communication to other users.

Any other comments?
None.

References

[1] ARORA, A., YANG, X., JHENG, S. Y., AND DELL, M. Linking representations with multimodal contrastive learning. *arXiv preprint arXiv:2304.03464* (2023).

[2] BERT, S. stackexchange, 2021.

[3] BOWMAN, S. R., ANGELI, G., POTTS, C., AND MANNING, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[4] CHEN, X., FANG, H., LIN, T.-Y., VEDANTAM, R., GUPTA, S., DOLLÁR, P., AND ZITNICK, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[5] CRASWELL, N., MITRA, B., YILMAZ, E., CAMPOS, D., AND LIN, J. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021), pp. 1566–1576.

[6] DUNN, M., SAGUN, L., HIGGINS, M., GUNEY, V. U., CIRIK, V., AND CHO, K. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179* (2017).

[7] ETHAYARAJH, K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512* (2019).

[8] FADER, A., ZETTLEMOYER, L., AND ETZIONI, O. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), pp. 1156–1165.

[9] FALKNER, S., KLEIN, A., AND HUTTER, F. BOHB: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning* (10–15 Jul 2018), J. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 1437–1446.

[10] GEBRU, T., MORGENSTERN, J., VECCHIONE, B., VAUGHAN, J. W., WALLACH, H., AU2, H. D. I., AND CRAWFORD, K. Datasheets for datasets, 2021.

[11] GUARNERI, J. *Newsprint Metropolis*. University of Chicago Press, 2017.

[12] HADSELL, R., CHOPRA, S., AND LECUN, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 2, IEEE, pp. 1735–1742.

[13] HE, K., GKIOXARI, G., DOLLÁR, P., AND GIRSHICK, R. Mask r-cnn. *Proceedings of the IEEE international conference on computer vision* (2017), 2961–2969.

[14] HENDERSON, M., BUDZIANOWSKI, P., CASANUEVA, I., COOPE, S., GERZ, D., KUMAR, G., MRKŠIĆ, N., SPITHOURAKIS, G., SU, P.-H., VULIĆ, I., ET AL. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI* (2019), pp. 1–10.

[15] IYER, S., DANDEKAR, N., AND CSERNAI, K. First quora dataset release: Question pairs, 2017.

[16] JOHNSON, J., DOUZE, M., AND JÉGOU, H. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data 7*, 3 (2019), 535–547.

[17] KHASHABI, D., NG, A., KHOT, T., SABHARWAL, A., HAJISHIRZI, H., AND CALLISON-BURCH, C. Gooaq: Open question answering with diverse answer types. *arXiv preprint arXiv:2104.08727* (2021).

[18] KWIATKOWSKI, T., PALOMAKI, J., REDFIELD, O., COLLINS, M., PARIKH, A., ALBERTI, C., EPSTEIN, D., POLOSUKHIN, I., DEVLIN, J., LEE, K., ET AL. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics 7* (2019), 453–466.

[19] LEWIS, P., WU, Y., LIU, L., MINERVINI, P., KÜTTLER, H., PIKTUS, A., STENETORP, P., AND RIEDEL, S. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics 9* (2021), 1098–1115.

[20] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[21] OCKERBLOOM, J. M. Everybody's library questions: Newspaper copyrights, notices, and renewals, 2019.

[22] RAJPURKAR, P., JIA, R., AND LIANG, P. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822* (2018).

[23] REIMERS, N., AND GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[24] SILCOCK, E., D'AMICO-WONG, L., YANG, J., AND DELL, M. Noise-robust de-duplication at scale. *International Conference on Learning Representations* (2023).

[25] SONG, K., TAN, X., QIN, T., LU, J., AND LIU, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems 33* (2020), 16857–16867.

[26] WANG, T., AND ISOLA, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *International Conference on Machine Learning 119* (2020), 9929–9939.

[27] WILLIAMS, A., NANGIA, N., AND BOWMAN, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).

[28] YOUNG, P., LAI, A., HODOSH, M., AND HOCKENMAIER, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2* (2014), 67–78.

[29] ZHANG, X., ZHAO, J., AND LECUN, Y. Character-level convolutional networks for text classification. *Advances in neural information processing systems 28* (2015).