smith_2017_evaluating_visual_representations_for_topic_understanding_and_their_effects_on_manually_generated_topic_labels

Year

2017

Author(s)

Smith, Alison and Lee, Tak Yeon and Poursabzi-Sangdeh, Forough and Boyd-Graber, Jordan and Elmqvist, Niklas and Findlater, Leah

Title

Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels

Venue

TACL

Topic labeling

Fully automated, Manual

Focus

Primary

Type of contribution

Established approach

Underlying technique

Manual labeling

TF-IDF vector cosine similarity

Topic labeling parameters

Nr of considered topic words: {5, 10, 20}

Label generation

Manual labeling

On Amazon Mechanical Turk, one set of users viewed a series of individual topic visualizations and provided a label to describe each topic, while a second set of users assessed the quality of those labels alongside automatically generated ones.

Better labels imply that the topic visualization provide users a more accurate interpretation (labeling) of the topic.

Phase I: Labeling

For each labeling task, users see a topic visualization, provide a short *label* (up to three words), then give a longer *sentence* to describe the topic, and finally use a five-point Likert scale to rate their confidence that the label and sentence represent the topic well.

We also track the time to perform the task. Figure 2 shows an example of a labeling task using the network graph visualization technique with ten words.

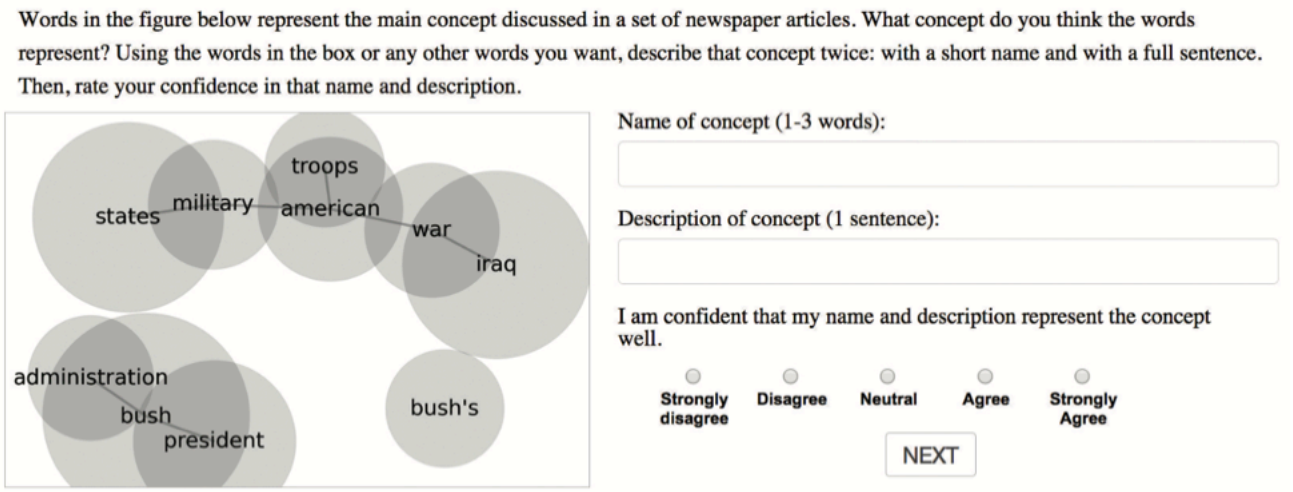

Figure 2: The labeling task for the network graph and ten words. Users create a short label and full sentence describing the topic and rate their confidence that the label and sentence represent the topic well.

For Phase I, we use a factorial design with factors of *Visualization* (levels: word list, word list with bars, word cloud, and network graph) and *Cardinal- ity* (levels: 5, 10, and 20), yielding twelve conditions. For each of the fifty topics in the model and each of the twelve conditions, at least five users per- form the labeling task, describing the topic with a label and sentence, resulting in a minimum of 3,000 label and sentence pairs.

Automatic labeling

Automatic labels are generated from representative Wikipedia article titles.

We first index Wikipedia using Apache Lucene. To label a topic, we query Wikipedia with the top twenty topic words to retrieve fifty articles. These articles' titles comprise our candidate set of labels. We then represent each article using its TF-IDF vector and calculate the centroid (average TF-IDF) of the retrieved articles.

To rank and choose the most representative of the set, we calculate the cosine similarity between the centroid TF-IDF vector and the TF-IDF vector of each of the articles. We choose the title of the article with the maximum cosine similarity to the centroid.

We do not include the topic words or Wikipedia title *n*-grams derived from our label set, as these labels are typically not the best candidates. Although other automatic labeling techniques exist, we choose this one as it is representative of general techniques.

Motivation

Unfortunately, manual labeling is slow and, while automatic labeling approaches exist, their effectiveness is not guaranteed for all tasks.

To better understand these problems, we use *labeling* to evaluate topic model *visualizations*.

These visualisations are:

- Word lists
- Word lists with bar graphs
- Word clouds
- Network graphs

Topic modeling

LDA

Topic modeling parameters

Nr of topics: 50

Alpha: 0.1

Beta: 0.01

Nr. of topics

50

Label

short *label* (up to three words) and longer *sentence* to describe the topic

Label selection

\

Label quality evaluation

The study includes two phases with different users. In Labeling (Phase I), users describe a topic given a specific visualization, and we measure speed and self-reported confidence in completing the task. In Validation (Phase II), users select the best and worst among a set of Phase I descriptions and an automatically generated description for how well they represent the original topics' documents.

Phase II: Validation
In the validation phase, a new set of users assesses the quality of the labels and sentences created in Phase I by evaluating them against documents associated with the given topic.
It is important to evaluate the topic labels in *context*; a label that superficially looks good is useless if it is not representative of the underlying documents in the corpus. Algorithmically generated labels (not sentences) are also included. Figure 3 shows an example of the validation task.

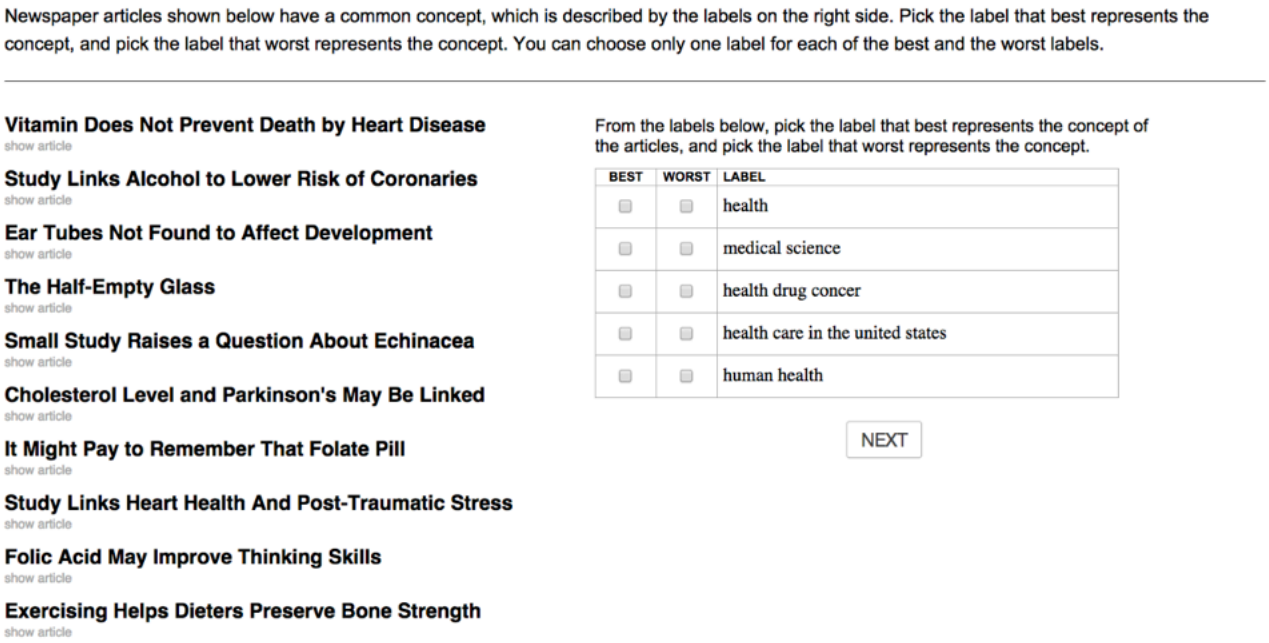

Figure 3: The validation task shows the titles of the top ten documents and five potential labels for a topic. Users are asked to pick the best and worst labels. Four labels were created by Phase I users after viewing different visualizations of the topic, while the fifth was generated by the algorithm. The labels are shown in random order.

The user-generated labels and sentences are evaluated separately.

For each task, the user sees the titles of the top ten documents associated with a topic and a randomized set of labels or sentences, one elicited from each of the four visualization techniques within a given cardinality.

The set of labels also includes an algorithmically generated label.

We ask the user to select the “best” and “worst” of the labels or sentences based on how well they describe the documents.

Documents are associated to topics based on the probability of the topic, z, given the document, d, $p(z|d)$.

Only the title of each document is initially shown to the user with an option to “show article” (or view the first 400 characters of the document).

All labels are lowercased to enforce uniformity. We merge identical labels so users do not see duplicates.

If a merged label receives a “best” or “worst” vote, the vote is split equally across all of the original instances (i.e., across multiple visualization techniques with that label).

Finally, we track task completion time.

Each user completes four randomly selected validation tasks, with the constraint that each task must be from a different topic. We also use ground truth seeding for quality control: each human intelligence tasks (HIT) includes one additional test task that has a purposefully bad label generated

by concatenating three random dictionary words. If the user does not pick the bad label as the “worst”, we discard all data in that HIT.

For Phase II, we compare descriptions across the four visualization techniques (and automatically generated labels), but only *within* a given cardinality level rather than *across* cardinalities.

Results

Perhaps unsurprisingly, there is no meaningful difference in the quality of the labels produced from the four visualization techniques.
However, simple visualizations (word list and word cloud) support a quick, first-glance understanding of topics, while more complex visualizations (network graph) take longer but reveal relationships between words.

Also, user-created labels are better received than algorithmically-generated labels, but more detailed analysis uncovers features specific to high-quality labels (e.g., tendency towards abstraction, inclusion of phrases) and the types of topics for which automatic labeling works. These findings motivate future automatic labeling algorithms.

Assessors

For validation in Phase II, we use the first five labels and sentences collected for each condition for a total of 3.000 labels and sentences. These are shown in sets of four (labels or sentences) dur- ing Phase II, yielding a total of 1,500 (3,000/4 + 3,000/4) tasks. Each HIT contains four validation tasks and one ground truth seeding task, for a total of 375 HITs. To increase robustness, we validate twice for a total of 750 HITs, without allowing any two labels or sentences to be compared twice.

Domain

Paper: Topic labeling

Dataset: News

Problem statement

We compare labels generated by users given four topic visualization techniques— word lists, word lists with bars, word clouds, and network graphs—against each other and against automatically generated labels. Our basis of comparison is participant ratings of how well labels describe documents from the topic. Our study has two phases: a labeling phase where participants label visualized topics and a validation phase where different participants select which labels best describe the topics' documents. Although all visualisations produce similar quality labels, simple visualizations such as word lists allow participants to quickly understand topics, while complex visualizations take longer but expose multi-word expressions that simpler visualiza- tions obscure. Automatic labels lag behind user-created labels, but our dataset of manually labeled topics highlights linguistic patterns (e.g., hypernyms, phrases) that can be used to improve automatic topic labeling algorithms.

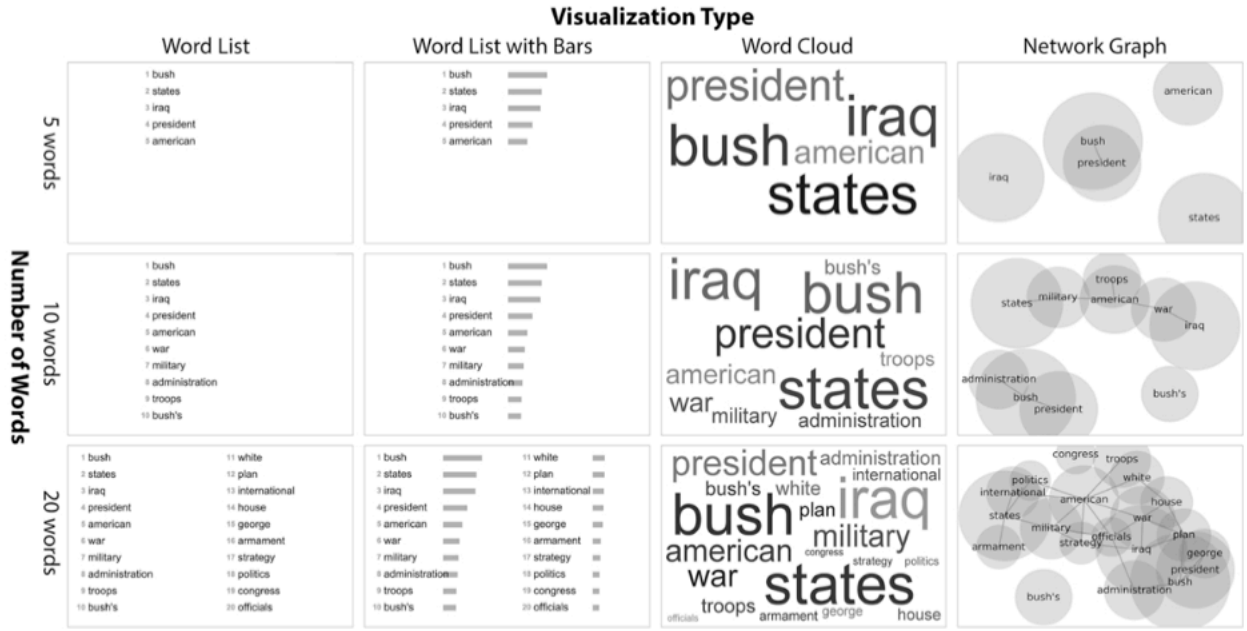

Figure 1: Examples of the twelve experimental conditions, each a different visualization of the same topic about the George W. Bush presidential administration and the Iraq War. Rows represent cardinality, or number of topic words shown (five, ten, twenty). Columns represent visualization techniques. For **word list** and **word list with bars**, topic words are ordered by their probability for the topic. **Word list with bars** also includes horizontal bars to represent topic-term probabilities. In the **word cloud**, words are randomly placed but are sized according to topic-term probabilities. The **network graph** uses a force-directed layout algorithm to co-locate words that frequently appear together in the corpus.

Corpus

Origin: New York Times

Nr. of documents: 7, 156

Details:

- articles from January 2007

Document

NYTimes article

Pre-processing

```
@article{smith_2017_evaluating_visual_representations_for_topic_understanding_and_their_effects_on_manually_generated_topic_labels,
    title = "Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels",
    author = "Smith, Alison  and
      Lee, Tak Yeon  and
      Poursabzi-Sangdeh, Forough  and
      Boyd-Graber, Jordan  and
      Elmqvist, Niklas  and
      Findlater, Leah",
    journal = "Transactions of the Association for Computational Linguistics",
    volume = "5",
    year = "2017",
    address = "Cambridge, MA",
    publisher = "MIT Press",
    url = "https://aclanthology.org/Q17-1001",
    doi = "10.1162/tacl_a_00042",
    pages = "1--16",
    abstract = "Probabilistic topic models are important tools for indexing, summarizing, and analyzing large document collections by their themes. However, promoting end-user understanding of topics remains an open research problem. We compare labels generated by users given four topic visualization techniques{---}word lists, word lists with bars, word clouds, and network graphs{---}against each other and against automatically generated labels. Our basis of comparison is participant ratings of how well labels describe documents from the topic. Our study has two phases: a labeling phase where participants label visualized topics and a validation phase where different participants select which labels best describe the topics{'} documents. Although all visualizations produce similar quality labels, simple visualizations such as word lists allow participants to quickly understand topics, while complex visualizations take longer but expose multi-word expressions that simpler visualizations obscure. Automatic labels lag behind user-created labels, but our dataset of manually
```

```
labeled topics highlights linguistic patterns (e.g., hypernyms, phrases) that
can be used to improve automatic topic labeling algorithms.",
}
```

#Thesis/Papers/BS