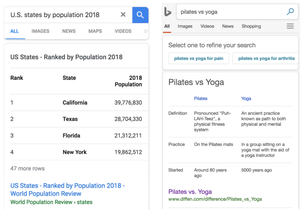
Generating Titles for Web Tables

Braden Hancock, Hongrae Lee, Cong Yu
bradenjh@stanford.edu, {hrlee, congyu}@google.com

Motivation

- Modern search engines no longer simply return links to relevant webpages. Often, the most relevant result in response to a user query has a semi-structured format.



- However, separating a table from its original context removes important clues that help a user interpret its contents and trust its relevance. Descriptive titles provide this crucial missing context.

Rocky Academy Award Nominations (1977)

Award	Result	Nominee
Best Picture	Won	Robert Chartoff and Irwin Winkler
Best Director	Won	John G. Avildsen
Best Actor	Nominated	Sylvester Stallone
Best Actress	Nominated	Talia Shire

6 more rows

vs

Award	Result	Nominee
Best Picture	Won	Robert Chartoff and Irwin Winkler
Best Director	Won	John G. Avildsen
Best Actor	Nominated	Sylvester Stallone
Best Actress	Nominated	Talia Shire

6 more rows

- We demonstrate how a model can be trained to automatically generate (and not merely extract) high-quality, descriptive titles for web tables using their surrounding metadata.

Dataset

- **Data:** To train and test our model, we collected a dataset of 10,102 web tables from 1384 domains, a subset of the tables returned as featured snippets to user queries on Google over a span of five months from January-May 2017.
- **Labels:** Each table in the dataset was shown (in context) to three crowdworkers, who proposed titles. If the same title was suggested by more than one crowdworker, use that one; otherwise, use the most descriptive title (as measured by word length).

90% The percent of crowdsourced titles containing at least one proper noun: proper handling of proper nouns is very important for title generation.

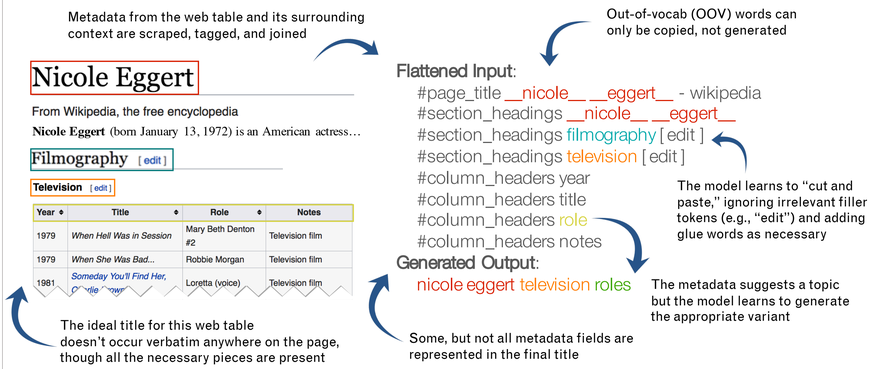
83% The percent of crowdsourced titles that were *composed* rather than *copied* (i.e., the title did not occur verbatim anywhere on the page).

45% The percent of crowdsourced titles containing Out-of-Vocab (OOV) words for a vocabulary of ~18k tokens from ~8k tables.

- **Failed Attempts:** Before paying crowdworkers to generate our training set, we attempted to build training sets using heuristics:
  - Heuristic 1: Use as titles the search queries that led users to the table
    - Problem: Many pages have more than one table; which one answered the query?
  - Heuristic 2: Use as titles the contents of the <caption> tag
    - Problem: Most captions make lousy titles (see Observations)

Main Idea

- **Problem:** The context required to fully understand a web table is often distributed among many different elements on the page. How should these pieces be found and stitched together to form a high-quality title?
- **Idea:** Train a seq2seq model to learn when to **copy** tokens from the table/page metadata and when to **generate** tokens from its vocabulary.
- **Result:** The copy mechanism helps the title stay relevant (e.g., using the appropriate rare entities) while the generation mechanism helps it stay readable (e.g., by connecting short nouns and phrases with the appropriate articles, prepositions, conjunctions, etc.)



Model

- **Inputs:** Scrape metadata from web tables and their surrounding context to try to capture all the relevant pieces of an appropriate title.

Field	Description
Page title*	Tokens inside the <title> tag nested in the <head> tag of the web page
Section headings*	Tokens in <h1>, <h2>, etc. tags with increasing priority, starting with the nearest
Table captions*	Captions inside <caption> tags
Spanning headers*	Headers in <th> tags that span all columns
Column headers*	Headers in <th> tags
Prefix text	Up to 200 tokens preceding the table until a new table or section boundary
Suffix text	Up to 200 tokens following the table until a new table or section boundary
Table rows	Text inside a <tr> tag, comma delimited with other tags (e.g., <td>) removed

- **Observation:** Interestingly, the actual contents of the table (other than header rows) contain very little relevant information for title generation.

2016 Olympic Medal Count

Rank	Country	Total Medals
1	United States	121
2	Great Britain	67
3	China	70
4	Russia	55
5	Germany	42

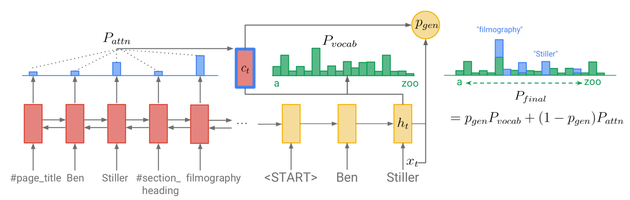
GDP by Country (Millions)

Rank	Country	GDP ($ Millions)
1	United States	18.6
2	China	11.2
3	Japan	4.9
4	Germany	3.5
5	United Kingdom	2.6

Global Average Internet Speeds

Rank	Country	Speed (Mb/s)
1	South Korea	26.7
2	Sweden	19.1
3	Norway	18.8
4	Japan	17.4
5	Netherlands	17.0

- **Architecture:** We use a pointer-generator network, a seq2seq model with a learned soft switch that dictates how much it favors words from the input vs words from its vocabulary.



Observations

- **Observation 1:** Most captions in the wild (text in a <caption> tag) make lousy titles. They are often either too verbose (e.g., a multi-sentence caption for an academic figure) or just one link in a long chain of relevant pieces of information, as shown below:

Page Title: 1936–37 NHL season
Section Heading: Regular Season
Section Heading: Final Standings
Caption: American Division

Title: 1936-37 NHL Regular Season American Division Final Standings

Page Title: The Beach at Anse Canot
Section Heading: Anse Canot
Section Heading: What's Nearby
Caption: Attractions

Title: The Beach at Anse Canot Nearby Attractions

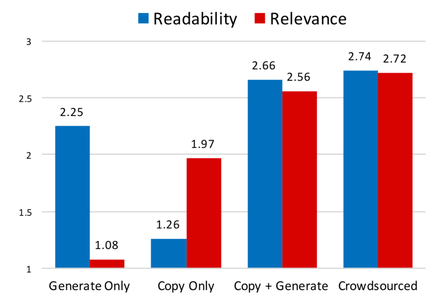
- **Observation 2:** An artifact of using token-wise log-likelihood as the loss function results in tokens often being repeated back-to-back.
  - For example, if the model is unsure whether the title will be "Highest Salaries of NBA Players" or "Highest Salaries in the NBA", it may generate the title "Highest Salaries NBA NBA NBA" to be sure to get credit for the "NBA" token.
  - Solution: Restrict the model to only use each token a maximum of once per title.
  - Result: Instant 4.5 ROUGE score boost.
    - 95% of our crowdsourced titles have no duplicate tokens to begin with.
    - Many valid titles with duplicate tokens can be paraphrased (e.g., "List of Mayors of Chicago" → "List of Chicago Mayors").

Results

- **Human Evaluation:** Evaluators assessed the readability and relevance of 200 titles on a held-out test split.

Model	Relevance	Readability	ROUGE-1	ROUGE-2	ROUGE-L
Page Title	2.25	2.41	0.510	0.369	0.461
Section Heading	2.29	2.56	0.476	0.315	0.411
Generate Only	1.08	2.25	0.168	0.064	0.151
Copy Only	1.97	1.26	0.384	0.221	0.240
Copy + Generate	**2.56**	**2.66**	**0.647**	**0.485**	**0.574**
Crowdsourced	2.72	2.74			

- **Baselines:** We explored using the page title or nearest section heading directly, as well as limiting our model to only copy or only generate. We also compared to human-generated titles.



- **Summary:** The generator alone is readable (good language model) but irrelevant, and the copier alone is relevant (good entity tagger) but unreadable. A model equipped with both does better on both metrics and approaches human performance.