Task Proposal: Abstractive Snippet Generation for Web Pages

Shahbaz Syed * **Wei-Fan Chen** † **Matthias Hagen** ** **Benno Stein** ‡
Henning Wachsmuth † **Martin Potthast** *
*Leipzig University †Paderborn University
**Martin-Luther-Universität Halle-Wittenberg ‡Bauhaus-Universität Weimar
<shahbaz.syed@uni-leipzig.de>

Abstract

We propose a shared task on abstractive snippet generation for web pages, a novel task of generating query-biased abstractive summaries for documents that are to be shown on a search results page. Conventional snippets are extractive in nature, which recently gave rise to copyright claims from news publishers as well as a new copyright legislation being passed in the European Union, limiting the fair use of web page contents for snippets. At the same time, abstractive summarization has matured considerably in recent years, potentially allowing for more personalization of snippets in the future. Taken together, these facts render further research into generating abstractive snippets both timely and promising.

1 Task Overview

The task of abstractive snippet generation can be defined as follows: Given a user query and a web page, generate an abstractive summary of the web page's content that conveys how it relates to the user's information need. A key aspect of this task is that the summary should be abstractive in nature, i.e., text reuse should be avoided as much as possible, whereas named entities and other facts that cannot be changed in terms of their phrasing should be retained. We perform both automatic as well as manual evaluation (via crowdsourcing) of the submitted models. To ensure reproducibility and blind evaluation, we employ the cloud-based evaluation platform TIRA (Potthast et al., 2019),[1] which facilitates software submission and blind evaluation on a hidden test set that is otherwise inaccessible. This ensures that participants cannot unwittingly optimize their approach against a test set. All data and code developed as part of the shared task will be publicly shared after its termination.

2 Motivation

A snippet on a search results page is a short summary accompanying each retrieved web page for a given user query. Conventionally, snippets are extractive summaries, where a few sentences containing the query's terms are extracted from the web page and arranged for display. With the query's terms highlighted in bold, one may decide at a glance whether a given web page merits further investigation with respect to one's information need. However, despite the relative ease with which snippets can be generated, and their near-universal use on modern web search engines, extractive snippets have limitations in terms of their expressiveness. More urgently, they have also increasingly become subject to copyright disputes and claims,[2] in particular from news publishers who lobbied for fair use, rendering extractive snippets from news almost or even entirely infeasible in some countries.

As a way forward, we envision *abstractive* snippets as an alternative (Potthast et al., 2018). An abstractive snippet is a query-biased abstractive summary of a web page that has minimal text reuse. Abstractive snippets have been shown to be on a par with extractive ones in terms of enabling search engine users to identify relevant results (Chen et al., 2018). Further, they are ideally suited to advance explainability and personalization of search results (Chen et al., 2020). Explanations may include reasons for ranking a page high or low, and personalization may hint either at information that the user has not seen elsewhere, or at adaptations to the user's experience level on a given subject. Owing to the recent advances in abstractive summarization (Lin and Ng, 2019) and text synthesis technology (Radford et al., 2019), we believe this is the right time to delve into abstractive

[1] TIRA, https://www.tira.io

[2] https://juliareda.eu/eu-copyright-reform/extra-copyright-for-news-sites/

Proceedings of The 13th International Conference on Natural Language Generation, pages 237–241,
Dublin, Ireland, 15-18 December, 2020. ©2020 Association for Computational Linguistics

Query: Treasury of Humor
Snippet: anchor context Asimov, on the other hand, proposes (in his first jokebook, Treasury of Humor) that the essence of humour is anticlimax: an abrupt change in point of view, in which trivial matters are suddenly elevated in importance above those that would normally be far more important.
Document [. . .] Treasury of Humor is unique in that in addition to being a working joke book, it is a treatise on the theory of humor, propounding Asimov's theory that the essence of humor is an abrupt, jarring change in emphasis and/or point of view, moving from the crucial to the trivial, and/or from the sublime to the ridiculous [. . .]

Table 1: Example of an anchor context as training snippet. The original anchor text is highlighted.

snippet generation. To the best of our knowledge, abstractive snippets are not yet adopted for commercial use. Through this shared task, we aim to investigate the capability of state-of-the-art summarization models to generate snippets that can be reliably presented to users of a search engine (commercial or otherwise).

Given the many relevant text generation tasks hosted at INLG, our task makes for a strong addition to this catalog, inviting participants not only from the NLP community but also from information retrieval. Building on a long and successful series of shared tasks in various domains,[3] including the recent *TL;DR summarization challenge* at INLG (Syed et al., 2019), we strive to provide a supportive infrastructure, rigorous evaluations, and to share useful insights with the research community.

3 Task Description

Participants may employ any abstractive summarization technology, with the aim of generating summaries (in the sense of snippets) that are query-biased. Among the first to study this task was Hasselqvist et al. (2017), who summarized news documents by using named entities as queries from the CNN/DailyMail dataset (Hermann et al., 2015). Recently, Chen et al. (2020) proposed a model employing bidirectional generation to induce the query bias, Xie et al. (2020) use conditional self-attention, Roitman et al. (2020) unsupervised learning, and Mao et al. (2020) lexically constrained decoding.

Query: Customer Respect Index
Snippet: DMOZ description The Customer Respect Group: An international research and consulting firm, publishes the Online Customer Respect Index (CRI) and provides industry and company-specific research and analysis to help companies increase sales and customer retention by improving how they treat their customers online.
Document [. . .] The Customer Respect Group has been a trusted source of online benchmark data and strategic insight since 2003. While much of our work is in financial services, we have worked across a variety of industries including telecommunications, education, government, and retail. [. . .]

Table 2: Example of a DMOZ description as training snippet. The original anchor text is highlighted.

3.1 Data

Similar to other text generation tasks, snippet generation requires a large dataset for training neural models. Accordingly, we have prepared the Webis Abstractive Snippet Corpus (Chen et al., 2020),[4] a novel and large-scale corpus for abstractive snippet generation. This corpus has been mined from the ClueWeb09, the ClueWeb12, and the DMOZ Open Directory Project, extracting more than 3.5 million examples of the form ⟨query, snippet, document⟩.

The *snippets* come from two sources: (1) anchor contexts, i.e., the text surrounding the anchor text of a hyperlink (linking to another web page) on a web page (see Table 1), and (2) descriptions from the Directory of Mozilla, DMOZ, the largest open directory project (see Table 2). The anchor contexts are used for distant supervision, based on the assumption that an author deciding to link to another's web page explains why the web page is linked, and what is found there, by summarizing its content. To ensure readable and meaningful snippet surrogates suited to the task, we filtered them in a multi-step pipeline. In case of DMOZ descriptions, they lend themselves directly: The directory contains human-written descriptions for web pages which serve as concise, abstractive snippets.

Finally, for each ⟨snippet, document⟩ tuple, we generated a matching *query* to which the document is relevant and the abstractive snippet surrogate is semantically related, at least marginally. To this end, we extracted noun phrases with the Stanford POS tagger (Toutanova et al., 2003), using only those phrases that occur in both the snippet and the

[3] https://webis.de/events.html?q=shared+task

[4] https://webis.de/data.html#webis-snippet-20

document in our examples. This ensured that the queries are relevant and distinct with regard to their context in the corresponding web page. At most three such queries per tuple were generated, with an average length of 2.4 words each. To allow for synergy, we additionally consider web pages that have been judged relevant to topics at the TREC Web, Session, and Task tracks.[5]

Crowdsourcing is employed to evaluate the quality of (1) the anchor contexts, (2) the generated queries, and, (3) the anchor contexts when used directly as query-biased snippets. The quality of the DMOZ descriptions was not evaluated, given their high a-priori quality. We selected 200 ⟨query, anchor context, document⟩ triples to be assessed. For all the three crowdsourcing studies, each task was done by five workers. The mean score for an anchor context was calculated based on the following annotation scheme: *very bad* gets score -2, *poor* score -1, *okay* score 1, and *very good* score 2.

In the study of anchor context quality, workers were shown individual anchor contexts to validate that the anchor contexts remaining after our preprocessing steps are of high linguistic quality. On average, the quality score was 1.06, showing that the quality of the anchor contexts can be expected to be *okay*. Next, the annotators judged if the generated queries are important with respect to their respective anchor contexts to validate our query generation approach. The mean query quality score was 0.28, showing the overall query quality is just above average. Lastly, we studied if the anchor contexts can be used directly as query-biased snippets by showing the entire triple to the workers. Here, the average score was -0.08, underlining that the anchor contexts may allow for distantly supervised training, but not close supervision.

Altogether, the three crowdsourcing studies have given us confidence that the anchor contexts we mined are reasonably well-suited to serve as summaries of their linked web documents, and that the queries generated for them serve as a reasonable point of connection between them. By extension, this also applies to the DMOZ descriptions, since high writing quality can be presumed here. For additional details about the corpus, we refer readers to Chen et al. (2020).

Participants are free to split this dataset into training and validation sets accordingly. The test dataset comprises 500 examples held out from the training data, each of which is manually inspected by multiple annotators to ensure quality. These are further divided into two subsets, one for automatic evaluation shared on the leaderboard, and one (truly hidden test set) for the final manual evaluation.

3.2 Protocol

We follow a similar protocol as the TL;DR challenge (Syed et al., 2018). The shared task is split into three phases: (1) Participants train suitable models using our dataset on their own hardware. (2) With the submission system open, participants deploy their models on TIRA and generate snippets for the provided test set. (3) After the submission deadline, the generated snippets from participants are manually evaluated via crowdsourcing. Ideally, Phase 1 begins three months before Phase 2 to ensure sufficient time for training. During Phase 2, the deployed models are automatically evaluated using multiple metrics to provide fast approximations of performance. All scores will be visible on a public leaderboard, with participants still being able to submit additional models at their discretion.

To ensure blind evaluation and reproducibility, the trained models are submitted as working software that generates snippets for a given set of ⟨query, document⟩ tuples. Participants deploy their software and all required dependencies on a virtual machine provided by the organizers. The test dataset is not accessible to participants while the competition is running; test set snippets are generated offline on the aforementioned virtual machine, without direct input from the participants. All evaluation runs are started from a clone of the participant's virtual machine, without network access, such that no test set data can be leaked. We operate the cloud infrastructure as well as the TIRA evaluation platform ourselves, so that no third party needs to be involved.

We plan the following schedule for the abstractive snippet generation task:

- **December 15th, 2020.** The shared task is announced along with the training data.
- **February 15th, 2021.** The submission system and public leaderboard are open. Participants can deploy and test models on the automatic evaluation test set.
- **May 15th, 2021.** This is the deadline for software submission; manual evaluation begins.

[5] TREC, https://trec.nist.gov

We estimate up to three weeks for completing the manual evaluation and their presentation on the public leaderboard. The shared task's findings are then presented at the following INLG, as was done for the TL;DR challenge (Syed et al., 2019).

3.3 Evaluation

Our evaluation is based on that of Chen et al. (2020) as a two-step process involving intrinsic and extrinsic evaluation. The intrinsic evaluation assesses multiple properties of a snippet: *text reuse*, *faithfulness* (no hallucinations), and *fluency*. Extrinsic evaluation assesses their *adequacy* in the context of being used within a search engine. A combination of relevant automatic metrics and manual evaluation are used in both the scenarios. While results of the automatic metrics are shared on the leaderboard throughout the duration of the task, those of the manual evaluation will be shared later. This is primarily due to the cost constraint of selecting only the top-performing models on the automatic metrics for the subsequent manual evaluation via crowdsourcing.

Intrinsic Evaluation For an overall comparison of the generated snippets to the ground truth, we employ the n-gram-based ROUGE as well as the contextual embedding-based BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019). BERTScore computes similarity by aligning the generated and the reference snippet on a token level with the objective of maximizing the cosine similarity between their contextual embeddings. MoverScore measures the semantic distance between the two by using Word Mover's Distance (Kusner et al., 2015) operating over the n-gram embeddings pooled from their BERT representations. This combination of metrics provides a decent approximation of the model's performance on both lexical and semantic levels. For assessing *text reuse*, we use the ROUGE-L precision score between the generated snippet and its source document. A lower precision implies less text reuse.

Evaluating *faithfulness* is done by calculating the ratio of noun phrases preserved by the generated snippet for a given document: $|S \cap \hat{S}|/|\hat{S}|$, where S is the set of noun phrases in a document, and $\hat{S}$ is the set of noun phrases in its generated snippet. Here, a noun phrase is defined with the restriction of being a head noun and an optional adjective, which have also been exclusively considered for query generation. This ratio approximates the amount of content units from the document that are preserved by the generated snippet.

Finally, *fluency* is judged manually via crowdsourcing (for the top-performing models) where (up to five) workers score a snippet's fluency on a 4-point Likert scale from *very bad* via *bad* and *good* to *very good*. Initially, however, fluency is indicated on the leaderboard as the perplexity of the generated snippet derived from a state-of-the-art language model as done by Chen et al. (2020).

Extrinsic Evaluation We assume that an *adequate* snippet of a web page summarizes its content in a query-biased manner and helps users identify relevant documents to the query from a given list of search results, where only the snippets are presented for each document. To this end, we set up a crowdsourcing experiment which simulates a typical search scenario. Our hidden test set contains topics from TREC tracks that have at least three relevant and three irrelevant documents judged in their corresponding datasets. Participating models generate snippets for a given topic (query) and its six documents with relevance judgments. Human annotators then judge each snippet with respect to its relevance to the given search query. We envision using up to 50 topics for the extrinsic evaluation.

4 Conclusion

We believe that our shared task will open new avenues to study abstractive snippet generation and query-biased summarization in general. By analyzing the performance of existing abstractive summarization technology from various perspectives, carrying out a comprehensive qualitative evaluation, and openly publishing all our data, code, and findings, we intend to make a meaningful contribution to the community in constrained text generation.

Abstractive summarization may hold the key to future web search technology, where a search engine not only explains to its users how a given web page is relevant to their current information need, but also why a given web page might be particularly relevant to them, personally. Although, even with current technology, we are still far removed from pursuing this goal, enabling contrained abstractive summarization is the first into this direction. Moreover, since web search engines currently operate perhaps the largest deployments of summarization technology, it is vitally important for our information society's ecosystem to maintain the ability to generate snippets in a copyright-compliant way.

References

Wei-Fan Chen, Matthias Hagen, Benno Stein, and Martin Potthast. 2018. A User Study on Snippet Generation: Text Reuse vs. Paraphrases. In *41st International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1033–1036. ACM.

Wei-Fan Chen, Shahbaz Syed, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Abstractive Snippet Generation. In *The Web Conference (WWW)*. ACM.

Johan Hasselqvist, Niklas Helmertz, and Mikael Kågebäck. 2017. Query-based abstractive summarization using neural networks. *arXiv preprint arXiv:1712.06100*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems (NeurIPS)*, pages 1693–1701.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, pages 9815–9822. AAAI Press.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation.

Martin Potthast, Wei-Fan Chen, Matthias Hagen, and Benno Stein. 2018. A Plan for Ancillary Copyright: Original Snippets. In *2nd International Workshop on Recent Trends in News Information Retrieval (NewsIR 2018) at ECIR*, volume 2079 of *CEUR Workshop Proceedings*, pages 3–5.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In *Information Retrieval Evaluation in a Changing World*, The Information Retrieval Series. Springer, Berlin Heidelberg New York.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Haggai Roitman, Guy Feigenblat, Doron Cohen, Odelia Boni, and David Konopnicki. 2020. Unsupervised dual-cascade learning with pseudo-feedback distillation for query-focused extractive summarization. In *The Web Conference (WWW)*, pages 2577–2584.

Shahbaz Syed, Michael Völske, Nedim Lipka, Benno Stein, Hinrich Schütze, and Martin Potthast. 2019. Towards Summarization for Social Media - Results of the TL;DR Challenge. In *12th International Natural Language Generation Conference (INLG)*.

Shahbaz Syed, Michael Völske, Martin Potthast, Nedim Lipka, Benno Stein, and Hinrich Schütze. 2018. Task Proposal: The TL;DR Challenge. In *11th International Conference on Natural Language Generation (INLG)*, pages 318–321.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL/HLT 2003*, pages 173–180.

Yujia Xie, Tianyi Zhou, Yi Mao, and Weizhu Chen. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv:2002.07338*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 563–578. Association for Computational Linguistics.