Towards Summarization for Social Media

Results of the TL;DR Challenge

Shahbaz Syed[a] **Michael Völske**[b] **Nedim Lipka**[c] **Benno Stein**[b] **Hinrich Schütze**[d] **Martin Potthast**[a]

[a]Leipzig University [b]Bauhaus-Universität Weimar [c]Adobe Research [d]LMU Munich

Abstract

With most summarization research focused on the news domain and scientific papers, little is known about the capabilities of the state of the art at summarizing more informal text. Today, the vast majority of text on the web is informally written on social media, and staying on top of the fast-paced stream of posts originating from one's subscriptions and followees is a burden to many. The TL;DR challenge marks a first step towards developing new summarization technology for social media, focusing on abstractive summarization. This paper reports the results of the challenge and describes our manual evaluation of the submissions. Finally, we discuss the expected properties of a good summary after analyzing the comments provided by human annotators.

1 Introduction

Current research on abstractive summarization focuses primarily on the genre of news. This can be attributed to the ease of obtaining large amounts of news articles alongside suitable summary ground truth, greatly simplifying the corpus construction. However, the summaries found in the currently widely used corpora are either only highlights directly extracted from news articles, offering little abstraction and no coherent text, or headlines, which are short and not necessarily summaries, albeit occasionally abstractive. Furthermore, the common structure of news articles[1] introduces bias, since the lead paragraph usually already captures the most relevant information (Kedzie et al., 2018).

To foster the development of robust summarization technology, we need to venture off the beaten track and explore more diverse domains. In this regard, the recently published Webis-TLDR-17 corpus (Völske et al., 2017) provides for the first English summarization corpus from the domain of social media, consisting of 3 million posts alongside so-called TL;DR summaries.[2] The summaries found in this corpus are true summaries provided by the authors of a post, they often abstract over a subject matter, and they cover a much wider range of topics than generally found in news articles. Table 1 shows a comparison of the nature of ground truth summaries in the news and the social media domain. With permission from its creators, we used this corpus to organize the TL;DR challenge (Syed et al., 2018), inviting summarization researchers to test existing models as well as new ones. To ensure reproducibility as well as blind and semi-automatic evaluation, we adopted the cloud-based evaluation platform TIRA (Potthast et al., 2019). In addition to the automatic ROUGE metrics, we evaluate the submissions manually for summary effectiveness and text quality via crowdsourcing. In this paper, we report our findings, discuss what annotators consider when scoring summaries, and outline future directions for abstractive summarization research.

2 Related Work

Shared tasks on automatic summarization were first introduced at the Document Understanding Conferences (DUC).[3] In addition to new summarization technology, equal emphasis was given to formulating strong evaluation measures. Methods such as basic elements (Hovy et al., 2006), pyramid (Nenkova and Passonneau, 2004), and ROUGE (Lin, 2004) were introduced for automatically evaluating the content selection capabilities of the participating systems. Furthermore, Dang (2005) presented the first guideline for manually judging summary quality. In 2008, DUC became a summarization track at the Text Analysis Conference

[1] https://en.wikipedia.org/wiki/News_style

[2] TL;DR, short for "too long; didn't read", is a cliché reply bemoaning a post's excessive length, and has given rise to a practice of adding a summary at the end of long posts, introduced by that same letter sequence or variants thereof.

[3] https://www-nlpir.nist.gov/projects/duc/index.html

Example - CNN/DailyMail Corpus

Article

NASA will launch Space Shuttle Endeavour on February 7, which will be the first of five launches this year before the shuttle fleet is retired. Endeavour will blast off from the Kennedy Space Center in Florida on a 13-day mission to the international space station. The mission will include three spacewalks, NASA said. The shuttle will also deliver the final U.S. portion of the space station. This portion will provide more room for crew members. NASA plans to retire its space shuttles Discovery, Endeavour and Atlantis later this year. The space agency has been looking for places, such as museums, to house the shuttles after they are retired. Space Shuttle Discovery will be transferred to the Smithsonian National Air and Space Museum in Washington. The privilege of showing off a shuttle won't be cheap – about $29 million, NASA said.

Highlights

- This will be first of five launches this year before the shuttle fleet is retired
- NASA is scheduled to launch Space Shuttle Endeavour on February 7.
- Shuttle will deliver final U.S. portion of the international space station
- NASA has been looking for places to house the shuttles once they are retired

Example- Webis-TLDR-17 Corpus

Post

I'm so upset at myself. My boyfriend surprised me with an amazing, fancy dinner for our one year anniversary yesterday. I already wasn't feeling well when he told me we were going to dinner but when I saw what he planned I didn't have the heart to tell him I wasn't that hungry. In the end I pushed myself to eat the fixed menu he ordered for us and the bill was over 500, I couldn't handle it and after dessert I ended up going to the bathroom and throwing it all up.

I can't believe I wasted so much of his money and am so disappointed in myself for not speaking up and simply saying I didn't feel well. I feel like I've wasted the effort he put into planning this. I also feel like I missed out on some amazing food that we would usually never splurge for. He doesn't know I threw it up and I just told him I loved it because regardless of how I felt health wise I loved that he put in so much effort to make sure I felt special. But I can't stop stewing in my own feelings. Help.

TL;DR

my boyfriend is amazing and bought us an expensive anniversary dinner. Threw it all up, he doesn't know. Feel horrible guilt and FOMO

Table 1: Comparison of summary styles from the CNN/DailyMail and the Webis-TLDR-17 corpus. Emphasized text shows the extractive nature of the summary (highlights) for the news domain. The highlights are concatenated and used as the target summary for training summarization models. In contrast, the example from the Webis-TLDR-17 corpus exhibits higher abstraction, abbreviations and composition of multiple facts into single phrases.

(TAC)[4] with evaluation as an independent task (Automatically Evaluating Summaries of Peers, AESOP). Most of these efforts were limited to extractive summarization on comparably small datasets from specific domains, such as biomedical records, newswire articles, and opinions, since neural text generation had not yet become mainstream, rendering abstractive summarization much more difficult.

The first attempt at abstractive summarization was presented by Rush et al. (2015), which resulted in a subsequent surge in neural summarization research yielding promising results—we refer to Shi et al. (2018) for a comprehensive review. However, as most recent models have been evaluated exclusively on news corpora, our knowledge of their full capabilities is still superficial. Through the TL;DR challenge, we hope to close this gap.

[4]https://tac.nist.gov/

3 Survey of Submissions

Out of 16 registered participants, we received 5 submissions from 3 participants (2 from industry). In addition, we provided a `seq2seq-baseline` model with 2 layers, bi-LSTM, 256 hidden units and no attention. Participants trained models at their own premises and deployed them to a virtual machine on TIRA. Via TIRA's web interface, scripts were configured to generate summaries for a hidden test set and then remotely executed. Multiple runs were allowed for each participant.[5] Each run was fed to an automatic evaluator script to compute ROUGE scores. Each software and evaluator run on the test set was manually reviewed by organizers for errors and data leakage. After a successful review,

[5]Evaluating models on TIRA using ROUGE was allowed even after the submission deadline. Thus, a participant's technical paper may have a variation of the same model with different ROUGE scores, but was not manually evaluated.

the scores were shared on a public leaderboard.[6] Two participants provided their system descriptions. We did not receive any description for the `tldr-bottom-up` model.

Gehrmann et al. (2019) leveraged fine-tuned language models to generate abstractive summaries. They argue that excessive copying facilitated by the copy-attention mechanism hinders paraphrasing and information compression (abstraction). As part of the TL;DR challenge, they compared two summarization approaches (`pseudo-self-attn` and `transf-seq2seq`) demonstrating the effectiveness of transfer learning at generating abstractive summaries. Our manual evaluation confirms that these models generate concise and coherent summaries.

Tackling the same problem of excessive copying in pointer-generator models, Choi et al. (2019) proposed using Variational Autoencoder (VAE) in combination with an extractive summarization model. The `unified-pgn` model uses a BERT-based extractive model that is fine-tuned to select important sentences, which are then summarized using a pointer-generator network. In order to introduce diversity, the `unified-vae-pgn` model uses a VAE for generating summaries of the extracted important sentences. This multi-stage architecture preserves a substantial amount of key information while generating acceptable summaries as revealed in our manual evaluation. We refer readers to the system description papers for further details.

4 Evaluation

Summarization evaluation measures based on the n-gram overlap between an automatically generated summary and a ground truth summary give good approximations of a summarizer's content selection capabilities. They are thus widely adopted and the de facto standard for the evaluation of extractive summarization technology. For lack of better alternatives, however, these measures have also been directly applied to abstractive summarization, where higher n-gram overlap does not necessarily indicate higher quality in terms of abstraction. In general, the measures employed to date also fail to capture situations where generated summaries are better than the corresponding ground truth summaries, since the latter must be considered only one of many possible alternative summarizations. Therefore, besides computing a standard measure, we also carry out a series of manual assessments via crowdsourcing to evaluate both the sufficiency and the text quality of a generated abstractive summary. Below, after reviewing both approaches, we report on the results of the participating systems.

4.1 Automatic Evaluation

We begin with a novelty analysis as per See et al. (2017), calculating the fraction of n-grams in the summary that are absent from the text as its novelty (Table 2). The ground truth has the highest novelty, underlining the abstractive nature of self-authored summaries. Next, we used ROUGE (Lin, 2004) for automatic evaluation and report the F1-scores.[7] From Table 2 it is difficult to draw any conclusions just by looking at ROUGE scores. Furthermore, a key issue of ROUGE is that it does not provide any upper bounds for the quality of a summarization system (Schluter, 2017), thus warranting an extensive manual evaluation of the systems.

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
seq2seq-baseline	3	0	2	0.00	2.27	2.47	2.05	4.9
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

Table 2: ROUGE-1, 2, and L scores and novelty analysis for 1 to 4-grams of the generated summaries along with their average lengths in words.

4.2 Manual Evaluation

Using Amazon Mechanical Turk, we crowdsourced our manual evaluation within two tasks: preference scoring and quality scoring. One hundred randomly selected examples from the test set were scored in both tasks, where each HIT (Human Intelligence Task) was assigned to 3 workers. We employed master workers with a minimum approval rate of 95% and at least 10,000 approved HITs.[8]

Preference scoring. The DUC guidelines for manually evaluating summaries by Dang (2005) were designed for experts. Gillick and Liu (2010) reported that Mechanical Turk workers were unable

[6] https://www.tira.io/task/tldr-generation/

[7] https://github.com/pltrdy/rouge; we intentionally rounded off the scores in our evaluation script in order to show differences of at least one point on the ROUGE metric.

[8] We paid $0.80 per HIT for preference scoring and $0.20 for quality scoring at an average hourly rate of $8 and $825 total.

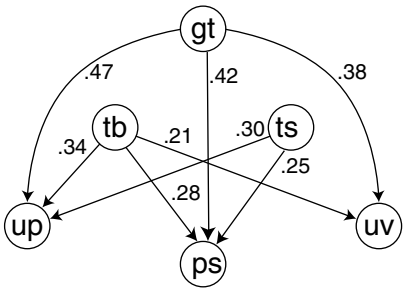

Figure 1: Summary of the preference scoring task: directed edges denote significantly higher scores ($p < 0.001$), and are annotated with effect sizes. The baseline model is much worse in comparison and hence not included. Key: **gt**: ground truth, **tb**: tldr-bottom-up, **ps**: pseudo-self-attn, **up**: unified-pgn, **ts**: transf-seq2-seq, **uv**: unified-vae-pgn.

Model	Sufficiency				Text quality			
	1	2	3	Avg.	1	2	3	Avg.
unified-pgn	2	38	60	2.11	6	68	26	1.78
unified-vae-pgn	4	30	66	2.13	9	62	29	1.78
transf-seq2seq	4	27	69	2.20	0	5	95	**2.70**
pseudo-self-attn	12	35	53	1.97	2	8	90	**2.67**
tldr-bottom-up	2	25	73	2.30	1	28	71	2.29
seq2seq-baseline	79	14	7	1.11	73	21	6	1.11
ground truth	2	8	90	**2.52**	0	15	85	2.57

Table 3: Sufficiency and text quality score distribution in the majority category.

to provide expert-like scores and had strong disagreements. Therefore, we kept the task as simple as possible: "Given a text and its summaries from all models (and the ground truth), score each summary for how well it summarizes the given text." We employed a four-point Likert scale ((1) very bad, (2) bad, (3) good, and (4) very good), since (Bishop, 1987) showed that presenting a middle alternative causes many people to choose it to escape uncertainty. Moreover, we asked for a written justification for each score. The scores collected reflect the summaries' overall quality, combining all aspects of summary quality relative to all other summaries, as perceived by the workers. Note that the summaries were shown in random order to prevent order effects. The score justifications required the workers to reflect about their judgments, and at the same time, they provide for an error analysis (see Section 4.3 for details). Moreover, the justifications allowed for double-checking whether workers actually read the summaries while scoring. Figure 1 shows which pairs of systems have significant differences along with effect sizes.[9]

Quality Scoring. Our second evaluation task was to independently assess a model's summaries across two specific qualitative dimensions. We adopt the term *sufficiency* to group multiple properties of a summary, such as informativeness, relevance, and focus. Similarly, *text quality* groups properties independent of the content, such as structure, coherence, grammar, and readability. In contrast to the first task, this gives workers specific goals and helps us to better differentiate between the models. Furthermore, it may help to identify if non-expert annotators can still produce reliable judgments without a guideline. Gillick and Liu (2010) cautioned that workers have difficulties distinguishing the content of a summary from its text quality. With that in mind, we devised two orthogonal three-level rating scales. With respect to *sufficiency*, workers could rate a summary as *insufficient* (incomplete and unrelated to the source text), as *barely acceptable* (missing the main point, but capturing relevant secondary information), or as *sufficient* (capturing the main point of the text). In terms of *text quality*, we distinguished the levels *badly written* (incoherent or major errors), *needs improvement* (minor errors breaking the flow, but understandable), and *well written* (no errors, coherent, and understandable).

Table 3 shows the score distribution for both dimensions in the majority category. For text quality, multiple models perform well compared to `ground truth`. Models with longer summaries (see Table 2), require further improvement in terms of text quality despite having a similar number of sufficient summaries. To compute significance, we assign the score of a summary to be the average of sufficiency and quality score. Figure 2 shows which pairs of systems had significant differences in scores along with effect sizes.

4.3 Error Analysis: Score Justifications

We manually reviewed all 2100 justifications given during the preference scoring task, and identified the summary aspects that most frequently influenced the scores. We further categorize these reasons under the two dimensions of sufficiency and text quality as shown in Table 4. These justifications may help the participants in improving their systems, and also aid the development of new models and evaluation methodologies. Moreover, com-

[9] We use Mann-Whitney U for pairwise comparison using Bonferroni correction.

Sufficiency	
Missing context (MC)	The summary does not provide any context, misses primary information or captures only secondary information.
Wrong sentiment (WS)	The overall sentiment of the post is either flipped or neutralized due to wrong negations.
Factually incorrect (FI)	Entities, such as names, locations, dates are wrongly reproduced, making the summary factually incorrect.
Overly simplistic (OS)	Summary lacks reasoning and necessary details making it too generic.
Text quality	
Bad grammar (BG)	A bad summary contains incorrect punctuations, wrong connectives, or formatting errors.
Incoherence (IC)	Improper flow of text which renders the summary meaningless.
Repetition (RP)	Excessive repetition of tokens.
Bad continuity (BC)	Summary starts off well but later culminates to gibberish text.

Table 4: Categories of worker criticism; the score of a summary was in many cases influenced by a combination of these aspects.

Model	**Sufficiency**				**Text quality**				**Pos.**
	MC	**WS**	**FI**	**OS**	**BG**	**IC**	**RP**	**BC**	
unified-pgn	94	12	11	6	40	81	22	10	56
unified-vae-pgn	52	6	21	9	39	61	12	8	100
transf-seq2seq	102	5	15	23	2	23	1	–	128
pseudo-self-attn	106	15	38	29	1	28	4	–	83
tldr-bottom-up	61	1	25	6	20	43	1	7	137
seq2seq-baseline	–	–	–	–	–	221	68	–	0
ground truth	69	1	11	14	10	12	0	3	178

Table 5: Distribution of summary aspects obtained from error analysis. The last column (positive) is the number of judgments (out of 300) where workers found no major problems with the summary .

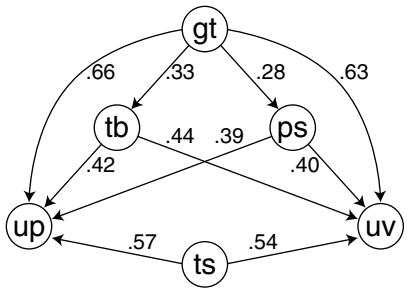

Figure 2: Summary of the quality scoring task: directed edges denote significantly higher scores ($p < 0.001$), and are annotated with effect sizes. The baseline model is much worse in comparison and hence not included. Key: **gt**: ground truth, **tb**: tldr-bottom-up, **ps**: pseudo-self-attn, **up**: unified-pgn, **ts**: transf-seq2-seq, **uv**: unified-vae-pgn.

paring the ordering of systems in Figure 1 and Figure 2, we see that master workers could differentiate the systems reasonably well without a guideline in the preference scoring task.

Table 5 shows the distribution of summary aspects for each model. *Missing context (MC)* was a key concern across all models where summaries failed to either capture enough details, or provide a proper reasoning, rendering them as *partial summaries* instead. This was prominent in the `transf-seq2seq` and `pseudo-self-attn` models, which produce shorter summaries that either lack relevant details or are overly simplistic (*OS*). However, these models generate the most coherent and readable summaries with very few cases of incoherence (*IC*) and no repetition (*RP*), obtaining an overall positive feedback. In contrast, the `tldr-bottom-up` and `unified-vae-pgn` models with much longer summaries preserved more information, but with issues in grammar (*BG*) or continuity (*BC*), leading to higher numbers of incoherent summaries.

5 Conclusion

Both `transf-seq2seq` and `pseudo-self-attn` generated the highest-quality text, but especially the latter often lacked information; `tldr-bottom-up` generated the most informative summaries (with acceptable text quality), followed by `transf-seq2-seq`. We found that, in the absence of a guideline, master workers provided reliable judgments by identifying influential summary aspects as seen in Table 4. All models struggled with capturing sufficient context spread throughout the posts, further aggravated by the casual writing style. Nevertheless, we observed encouraging results in terms of text quality. We envision that summarization will benefit from including formalisms of importance, argumentation, and reasoning into the models, while striking a balance between summary length and text quality. In the next edition of this task, we will foster corresponding contributions.

References

George F Bishop. 1987. Experiments with the middle response alternative in survey questions. *Public Opinion Quarterly*, 51(2):220–232.

Hyungtak Choi, Lohith Ravuru, Tomasz Dryjanski, Sunghan Rye, Donghyun Lee, Hojung Lee, and Inchul Hwang. 2019. Vae-pgn based abstractive model in multi-stage architecture for text summarization. In *TL;DR Challenge System Descriptions*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Sebastian Gehrmann, Zachary Ziegler, and Alexander Rush. 2019. Generating abstractive summaries with finetuned language models. In *TL;DR Challenge System Descriptions*.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*, pages 148–151.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006)*, pages 604–611. Citeseer.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 41–45.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2018. Neural abstractive text summarization with sequence-to-sequence models. *CoRR*, abs/1812.02303.

Shahbaz Syed, Michael Völske, Martin Potthast, Nedim Lipka, Benno Stein, and Hinrich Schütze. 2018. Task proposal: The tl;dr challenge. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 318–321.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.