Task Proposal - The TL;DR challenge

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

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

Abstract

The TL;DR challenge fosters research in abstractive summarization of informal text, the largest and fastest-growing source of textual data on the web, which has been overlooked by summarization research so far. The challenge owes its name to the frequent practice of social media users to supplement long posts with a "TL;DR"—for "too long; didn't read"—followed by a short summary as a courtesy to those who would otherwise reply with the exact same abbreviation to indicate they did not care to read a post for its apparent length. Posts featuring TL;DR summaries form an excellent ground truth for summarization, and by tapping into this resource for the first time, we have mined millions of training examples from social media, opening the door to all kinds of generative models.

1 Task overview

The task of participants is straightforward: given a social media posting, generate a summary. Ours is to evaluate the participants' approaches quantitatively and qualitatively, which will include measures from the literature as well as manual review via crowdsourcing. For diversity, we intend to offer additional evaluation categories that further constrain a summary, e.g., summaries that include the topic of an underlying discussion, summaries in the form of questions, or, for fun sake, wacky summaries. To ensure reproducibility, we employ the cloud-based TIRA[1] evaluation platform, to which participants will deploy working prototypes of their summarizers for blind and semi-automated evaluation.

[1] www.tira.io

2 Motivation

Text summarization ranks among the oldest synthesis tasks of computer science, addressing the nowadays almost stereotypical problem of information overload. Traditionally, the task has been tackled within natural language processing and information retrieval by extracting phrases from a to-be-summarized text. However, the task draws increasing attention from the machine learning community. Owing to advances in theory, algorithms, and hardware, the training of complex models has become feasible that *abstract* over the to-be-summarized text. Here, deep generative models have delivered some impressive results (Chopra et al., 2016; Nallapati et al., 2016; See et al., 2017; Chen and Bansal, 2018). Since these models need substantially large amounts of training data in order to understand and generate natural language text, the availability of suitable corpora is important. The most commonly used datasets for abstractive summarization, namely the Gigaword corpus (Graff and Cieri, 2003) and the CNN Dailymail news dataset (Hermann et al., 2015), comprise short articles from the news domain, representing only one of the many genres of written text. Target summaries in both these corpora are extractive where either the first sentence or some key points are combined together to train the model. In particular, there have been no resources covering user-generated content until now, severely limiting the range of applications of summarization research and technology on the web.

Social media platforms and search engines alike rely on showcasing contents to their users in order to maintain and increase user engagement. Besides the intelligent, personalized recommendation and retrieval systems which retrieve the right content at the right time, they mostly resort to extractive summarization techniques for presentation. While con-

tents with a high production value, such as news, regularly come with pre-produced summaries tailored to the most important platforms and search engines, this is not the case for user-generated content, which makes up for the vast majority of content on many platforms. At the same time, user-generated content, due to the informal nature of its writing, does not readily lend itself to extractive summarization techniques. Abstractive approaches may offer great value to these platforms by capturing the gist of a piece of user-generated content while harmonizing the style of presentation. This is particularly important in cases where the user interface is limited, such as that of mobile devices or the narrow audio-only interface of conversational AI agents.

Through our competition, we want to spur the development of novel model architectures and optimizations for generative models, in order to bridge the gap between the quality of automatic summaries and those written by humans. In accordance with the motto of INLG, we encourage discussions about a multitude of unsolved research questions related to text summarization and its evaluation. To the best of our knowledge, we are the first to tackle abstractive summarization of user-generated content at scale. As organizers of many previous shared tasks in the areas of natural language processing and machine learning, we look back on years of experience in provisioning and administering the infrastructure required.

3 Task Description

The task of participants is to provide a software that, given a text, generates an abstractive summary for it. This task is at the heart of modern summarization technology that may be used in the aforementioned scenarios. Participants are encouraged to perform any preprocessing/normalization of the provided training data as well as to incorporate third party data as they see fit. Their training process must accordingly be described in detail in the paper accompanying their final submissions.

The task is challenging, but—given recent advances in deep generative modeling—not impossible, anymore. Key to solving this task is to identify means that will allow for generating summaries that are short as well as self-explanatory, to work around the idiosyncratic usage in user-generated content, and to find levers to adjust summary quality. In this regard, a promising direction may also be the combination of traditional, extractive summarization approaches with deep generative models as demonstrated by Liu et al. (Liu et al., 2018).

Table 1: Sample content-summary pair

Content: not necessarily my lucky day , but some kids this is how it went was sitting out on the dock at a local lake with a friend sharing some beers . little boy aged 2-3 yrs old walks up with a wooden stick and starts poking at the water . it was windy out and the dock was moving , and sure enough the kid leans over just enough to topple head first into the water . i had already pulled my phone out and wallet out just in case i was to accidentally fall in so i went straight over and hopped in . saw his little hand reaching up and tossed him straight back onto the dock . walked him to his dad who didn ' t speak any english and was very confused why i had his son soaking wet . left later that day and saw the kid back on the dock ! it blew my mind.
TL;DR: saved a 2 year old from drowning at a lake because i was drinking beers with a friend .

3.1 Data

The competition can be immediately started, since its training dataset is already available (Völske et al., 2017). We have mined Reddit for user postings that include a TL;DR summary, collecting a total of 4,044,501 content-summary pairs (see Table 1 for an example). In terms of size, our dataset matches the state-of-the-art Gigaword corpus. This dataset covers a wide range of everyday topics, drawing examples from 32,778 *subreddits*, each of which focuses on a particular topic. The mining process was carefully adjusted to include and extract the many syntactical variants of TL;DR summaries while excluding automatically generated postings and summaries by bots. To ensure high quality, we frequently reviewed large samples of the data, adjusting the mining process until at least 95% of the samples were of sufficient quality. Each item of the dataset is a pair of posting and summary written by the same author.

For the competition, we will use a subset of this dataset comprising 3,084,410 of the content-summary pairs to harmonize the length distributions. In this subset, the average length of a posting is 211 words, ranging from from 100 to 400 words, and that of a summary is 25 words, ranging from 10 to 200 words. Participants are free to split this into training and validation sets as deemed fit. The test dataset comprises 1000 items held out from the training data, each of which has been carefully reviewed by at least three human annotators for quality. Of these, 800 will be used in the initial automatic evaluation runs, and the remaining 200 in a final round of manual evaluation.

3.2 Protocol

The competition will comprise three phases: (1) participants will train summarization models using

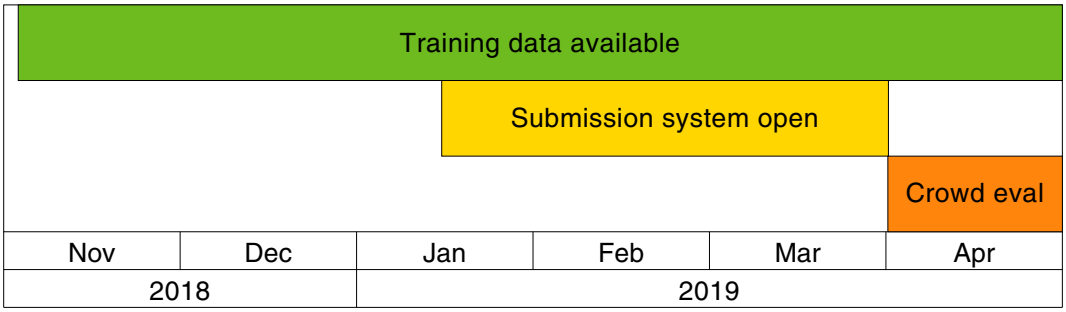

Figure 1: Planned timeline for the TL;DR challenge.

the provided training data on their own hardware; (2) with the submission system open, participants will deploy their trained models on the TIRA infrastructure; the systems will generate candidate summaries against the automatic evaluation test samples, without network access or direct involvement of the participants. (3) once the submission dates conclude, the candidate summaries generated by the submitted models will be evaluated by crowdsourcing workers against the private test set.

Phase (1) will begin two and a half months before Phase (2), from which point both will run in parallel until the submissions system closes. Participants will be able to train and submit models at their discretion. For the duration of Phase (2), submitted summaries will be automatically evaluated using the ROUGE measure against half of the test set samples. Automatic evaluation scores will populate the public leaderboard. The final portion of the test set will only be used in the manual evaluation round, so that overfitting against the leaderboard scores can be avoided.

To ensure blind evaluation, and reproducibility, the trained summarization models will be submitted as working *software* that performs summary inference given a set of input texts. Participants will deploy this software and all required dependencies on a virtual machine provided by the challenge organizers. The test dataset will not be accessible to participants while the competition is running; test set summaries will be generated offline on the aforementioned virtual machine, without direct input from the participants. All evaluation runs will be 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 need be involved. TIRA has been successfully employed for various large-scale competitions since 2012; it is battle-tested.

Finally, we will encourage participants to share their code bases in a central organization at GitHub as a kind of Open Source Proceedings.

3.3 Schedule

We envision the milestones shown in Figure 1 for scheduling the TL;DR challenge:

- **November 5th, 2018:** Challenge announced; training data available.
- **January 15th, 2019:** Submissions system and public leaderboard open. Challenge participants will be able to submit working summarizers to the TIRA infrastructure; the online leaderboard will be continuously updated to reflect the latest performance on the test set.
- **April 1st, 2019:** Deadline for final software submissions; crowd evaluation begins.

3.4 Evaluation

To determine the winners of the TL;DR challenge, we will deploy a two step process involving both automatic measures and a thorough human evaluation of the generated summaries. Content selection evaluation metrics such as ROUGE, BLEU, and METEOR, will be reported to provide participants with a first impression of the coherence and information capturing capabilities of their models. Additionally, embedding based metrics such as cosine similarity of word and sentence representations of the generated summaries will be reported against the reference summaries.

For qualitative evaluation, human annotators recruited via Amazon Mechanical Turk will read the candidate summaries and rate them based on the standard summary evaluation criteria established by the DUC competitions (Dang, 2005). Each generated summary will be judged by at least three annotators to ensure accuracy; annotators will rate individual summaries, as well as pairs of summaries

from different participants to establish preference and break ties. The final ranking, and the winner of the TL;DR challenge, will be derived from the human annotators' quality judgments. The crowdsourcing evaluation phase will employ 200 test samples not used during the automatic evaluation phase. Based on the number of submitted summarization systems, participation in the crowd evaluation phase may be limited to the top performers on the automatic evaluation leaderboard—based on our projections, up to approximately thirty submissions will be considered for crowd evaluation.

Some time after the conclusion of the competition, all testing data and annotator decisions will be made available to the research community at large; we expect that the analysis of the resulting data, and how it correlates with automatically computed ROUGE scores, will benefit the development of better evaluation metrics.

Outside of the ranking, we intend to offer evaluation scenarios in constrained summarization, such as generating summaries that include the topic of the underlying discussion, summaries in the form of questions, or wacky summaries deliberately including off-color vocabulary. We envision such scenarios to gain interest as summarization technology becomes integrated into conversational agents.

4 Conclusion

We strongly believe that our shared task proposal will encourage creation of diverse datasets for neural summarization. The TL;DR dataset poses a different set of challenges for neural generation models compared to News corpora. By emphasizing on the effectiveness and limitations of exisiting models through our challenge, the NLG community can focus on novel models and evaluation measures for developing better summarization technology.

References

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. pages 93–98. http://aclweb.org/anthology/N/N16/N16-1012.pdf.

Hoa T. Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*.

David Graff and Christopher Cieri. 2003. English Gigaword LDC2003T05. Web Download. Philadelphia: Linguistic Data Consortium.

Karl Moritz Hermann, Tomáš Kočiský, 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 (NIPS)*. http://arxiv.org/abs/1506.03340.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *CoRR* abs/1801.10198. http://arxiv.org/abs/1801.10198.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*. pages 280–290. http://aclweb.org/anthology/K/K16/K16-1028.pdf.

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. https://doi.org/10.18653/v1/P17-1099.

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, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*. pages 59–63. https://aclanthology.info/papers/W17-4508/w17-4508.