Fact-based Content Weighting for Evaluating Abstractive Summarisation

Xinnuo Xu[†], **Ondřej Dušek**[‡], **Jingyi Li**[†], **Verena Rieser**[†] **and Ioannis Konstas**[†]
[†]The Interaction Lab, MACS, Heriot-Watt University, Edinburgh, UK
[‡]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
xx6, jl125, v.t.rieser, i.konstas@hw.ac.uk
odusek@ufal.mff.cuni.cz

Abstract

Abstractive summarisation is notoriously hard to evaluate since standard word-overlap-based metrics are biased towards specific words in the human reference. We introduce a new evaluation metric which abstracts away from the word-level and instead is based on fact-level content weighting, i.e. relating the facts of the document to the facts of the summary. We follow the assumption that a good summary will reflect all relevant facts, i.e. the ones present in the ground truth (human-generated reference summary). We confirm this hypothesis by showing that our weightings are highly correlated to human perception and compare favourably to the recent manual highlight-based metric of Hardy et al. (2019).

1 Introduction

Text summarisation compresses long textual documents into short summaries while retaining the most important information from the source. In contrast to extractive summarisation, which directly copies the most relevant fragments, abstractive summarization retains the most important facts and expresses them via paraphrasing, aggregating and even inferring new facts. Recent advances in neural decoders led to a number of single-document summarisation systems that exhibit some level of abstraction in their outputs, usually in the simplest form of paraphrasing (See et al. (2017); Narayan et al. (2018); Liu and Lapata (2019), *inter alia*).

Evaluating abstractive summarisation remains an open challenge (Schluter, 2017; Kryściński et al., 2019): First, decoders are amenable to pathogeniessuch as hallucination and/or omission of important information, which are hard to capture using existing evaluation metrics (Cao et al., 2018; Rohrbach et al., 2018; Dušek et al., 2020). Second, most datasets used for abstractive summarisation only contain a single reference summary, e.g. (Narayan et al., 2018; Völske et al., 2017), which most existing automatic metrics evaluate against, e.g. ROUGE using exact n-gram overlap (Lin, 2004), and thus tend to downvote paraphrases.

We propose a new evaluation metric based on content weighting, where we abstract away from the particular surface form of the target summary, but represent it as facts using Semantic Role Labelling (SRL). In this way, we aim to better capture the semantic correctness of a summary, i.e. be more sensitive to hallucinations and omissions.[1]

In particular, we weight the facts present in the source document according to the facts selected by a human-written summary. This *alignment* is conducted using contextual, rather than token-level, embeddings, e.g., BERT (Devlin et al., 2019). For evaluation, we measure whether an automatically generated summary is able to capture the same facts as the target. We also show that the computed weights correlate well with human perception. Our code is available at https://github.com/XinnuoXu/CorrFA_for_Summarizaion.

2 Related Work

The problem of reference bias has been addressed in several ways. First, metrics based on token-level or wider context embedding similarities which aim to better capture paraphrases but remain largely word-oriented, e.g. (Sun and Nenkova, 2019; Zhang et al., 2019; Zhao et al., 2019; Clark et al., 2019). Goodrich et al. (2019) come close to our approach by using entity and relation extraction, but their approach is limited to texts that lend themselves to be represented by RDF triples.

An alternative is manual evaluation against the source document. This entails selecting content either using domain experts, e.g., the PYRAMID method (Nenkova and Passonneau, 2004), factoids

[1]Note that we do not make any claims about fluency, which we assume is less of a problem for neural text generation.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5071–5081
July 5 - 10, 2020.

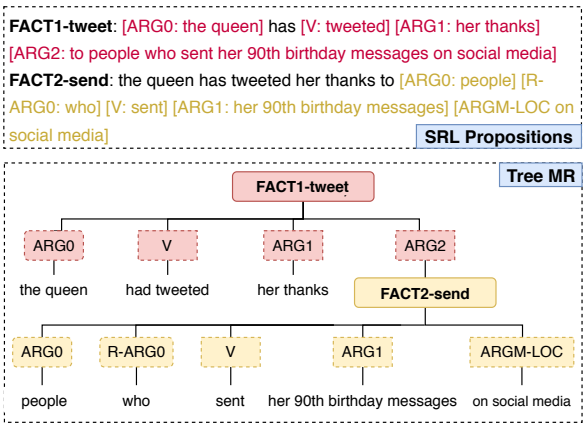

Figure 1: List of SRL propositions and corresponding tree MR with two facts for the sentence *"The queen has tweeted her thanks to people who sent her 90th birthday messages on social media"*.

(Teufel and van Halteren, 2004), or via crowdsourcing (Shapira et al., 2019; Hardy et al., 2019). However, evaluation based on a small human-labelled test set is noisy, time consuming, and costly. Xenouleas et al. (2019) propose a referenceless metric, which only checks properties of the summary, not its relation to the original document. Sun and Nenkova (2019) compare average token and sentence ELMo embeddings against the document and claim good (system-level) correlations.

Another option to avoid reference bias is question-based evaluation, either elicited manually (Clarke and Lapata, 2010; Narayan et al., 2018) or automatically (Scialom et al., 2019). However, it requires reference summaries as base for generating questions, thus only checking the summary contents indirectly.

3 Content Weighting

3.1 Fact Representation

We represent facts in a sentence by adapting SRL (Palmer et al., 2005), which roughly captures "who did what to whom" in terms of predicates and their arguments. Given a list of parsed propositions for a sentence,[2] each predicate-argument structure is considered as one separate *fact*, where the predicate stands for the event and its arguments are mapped to actors, recipients, time, place, etc (see Fig. 1). Following a simple observation that arguments can function as separate predicates themselves, we construct a hierarchical tree structure for the whole sentence. We create the tree meaning representation (MR) from the list of facts by choosing the fact with the largest coverage as the root and recursively build sub-trees by replacing arguments with their corresponding sub-facts (ARG2 in FACT1 is replaced by FACT2 in Fig. 1).[3]

3.2 Automatic Content Weighting

We compute argument and fact weights by measuring the similarity of facts/arguments in the original document and the target summary based on their BERT word embeddings (for content words only) and their distance in the tree MR. We denote tokens of a document D and its summary S as $\mathbf{t}^D = \{t_1^D, t_2^D, \cdots t_n^D\}$ and $\mathbf{t}^S = \{t_1^S, t_2^S, \cdots t_m^S\}$. To get their corresponding contextual embeddings e_k^D and e_k^S, we concatenate the two texts,[4] feed them into a pre-trained BERT model (Devlin et al., 2019) and take the contextualized embedding output from its last Transformer layer.

Argument-based weighting: We first represent the summary and the document as two sequences of leaf arguments[5] $\{A_1^D, A_2^D, \cdots A_N^D\}$ and $\{A_1^S, A_2^S, \cdots A_M^S\}$ respectively, and weight the i-th leaf argument in the document as:

$$w_i^a = \underset{j=1\ldots M}{\text{avg}} \text{cosdist}\left(E_i^D, E_j^S\right) \quad (1)$$

i.e. the average embedding cosine distance to all arguments in the summary. Argument embeddings E_i^D and E_j^S are average embeddings of content-word tokens belonging to the arguments:[6]

$$E_i^* = \underset{k \in A_i^*, k \notin \text{stops}}{\text{avg}} e_k^* \quad (2)$$

$* \in \{D, S\}$, "stops" denotes a list of stopwords.

Fact-based weighting: We can represent the summary and the document as two sequences of facts $\{F_1^D, F_2^D, \cdots F_{N'}^D\}$ and $\{F_1^S, F_2^S, \cdots F_{M'}^S\}$, and weight the i-th fact in the document by its average distance to facts in the summary:

$$w_i^f = \underset{j \in 1\ldots M'}{\text{avg}} d_{ij}^f \quad (3)$$

[2] We use the SRL implementation of He et al. (2018) found in https://allennlp.org with 86.49 test F1 on the Ontonotes 5.0 dataset.

[3] We avoid using sentence-level MRs such as AMR (Banarescu et al., 2013), since current state-of-the-art performance of parsers is far behind compared to the simpler SRL task.

[4] By concatenating, the information in each text can be embedded in each other through self-attention. This is useful since the summary sometimes contains additional and/or common-sense knowledge not captured in the document.

[5] For example, in Fig. 1, ARG0, V, ARG1 in FACT1, and all the arguments in FACT2 are leaf arguments in the sentence, whereas ARG2 in FACT1 is not.

[6] For example, in Fig. 1, "her" and "thanks" are two tokens directly attached to the argument ARG1 of FACT1. Thus, the embedding for ARG1 of FACT1 is the average embedding of these two tokens.

The fact-level distance d^f_{ij} is defined on top of argument weighting:

$$d^f_{ij} = \underset{A^D_l \in F^D_i, A^S_k \in F^S_j}{\text{avg}} \beta_{il}\beta_{jk} \lfloor \text{cosdist}\left(E^D_l, E^S_k\right) \rfloor_{>\gamma} \quad (4)$$

It is computed as the average cosine distance over embeddings of all leaf arguments in the subtrees of fact F^D_i in the document and fact F^S_j in the summary, which is (1) filtered by a threshold γ to discard argument pairs with weak semantic relation[7] and (2) weighted by MR tree distances of arguments to facts: $\beta_{il} = \frac{1}{\sqrt{\text{treedist}(F_i, A_l)}}$.[8]

4 Content-weighting-based Metrics

We now use these weights to introduce two metrics: **Corr-F** (fact-level) and **Corr-A** (argument-level). Let $\mathbf{w}^f_{gold}$ and $\mathbf{w}^f_{cand}$ denote the fact-level content weights calculated using the procedure from Section 3 based on human-reference and system-generated summaries, respectively. Similarly, $\mathbf{w}^a_{gold}$, and $\mathbf{w}^a_{cand}$ denote the argument-level weights. Corr-F is then the Pearson Correlation Coefficient (PCC) between $\mathbf{w}^f_{gold}$ and $\mathbf{w}^f_{cand}$. Corr-A is PCC between $\mathbf{w}^a_{gold}$ and $\mathbf{w}^a_{cand}$. In other words, Corr-F and Corr-A indicate whether the generated summary focuses on the informative main points in the document (i.e. the same points as the reference summary), on two different levels of granularity.

5 Metrics Evaluation

We validate our Corr-F and Corr-A metrics by collecting human judgements. In the following, we (1) collect content highlights from human judges using the Amazon Mechanical Turk platform[9] and calculate manual content weighting based on them, (2) calculate correlations of the manual content weights with our automatic content weights, (3) compare our metrics against existing reference-based ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019), as well as the referenceless manual HROUGE score (Hardy et al., 2019).[10]

We use the extreme summarisation dataset (XSum; Narayan et al., 2018), which consists of BBC articles and accompanying single-sentence summaries, i.e. sub-headlines of the original articles, professionally written by the authors of the articles. Due to the abstractive nature of the summaries, factoid content selection on phrase level is required beyond sentence-level extraction or token-level matching, making this dataset a popular test bed for abstractive summarisation.

We use the outputs of three recent abstractive summarization systems as evaluation targets for our metrics: (i) the Pointer-Generator model (PTGEN; See et al., 2017); (ii) the Topic-aware Convolutional Sequence-to-Sequence model (TCONVS2S; Narayan et al., 2018) and (iii) the abstractive summarization model using pretrained BERT encoders (BERTSUMABS; Liu and Lapata, 2019).[11]

5.1 Manual Annotation Collection

Manual Content Highlighting: By extending the framework of Hardy et al. (2019), we collect manual content highlights on fact and argument levels, where we present human judges with the source document and the gold summary, with one fact/argument typeset in bold. The judges are required to select phrases or sentences in the document that support the bolded fact/argument (see Figure 4-9 in Appendix B). In both cases, judges are allowed to select parts of the text with any granularity. We limit the number of allowed continuous chunks and the maximum number of words to encourage highlights of fact/argument level.[12] We employ 3 judges per document in both cases. We use the same 50 articles and gold summaries sampled from the XSum test set as Hardy et al. (2019).

Manual Content Weighting Calculation:
Argument Level: Given a document D and a summary S, we define the weight of each token t^D_k with respect to a summary argument A^S_j as:

$$w^t_{kj} = \frac{\text{NumH}\left(t^D_k, A^S_j\right)}{\text{NumA}\left(A^S_j\right)} \quad (5)$$

$\text{NumH}(t^D_k, A^S_j)$ denotes the number of times token t_k was selected and $\text{NumA}(A^S_j)$ is the total number of annotators who were shown A^S_j bolded. We use token weights to compute manual argument-level weights $\mathbf{w}^a_{man}$ (parallel to Eq. 1):

$$w^a_{man,i} = \underset{j=1\ldots M}{\text{avg}} \underset{t^D_k \in A^D_i}{\text{avg}} w^t_{kj} \quad (6)$$

[7] In this work, we set the threshold to 0.6.

[8] E.g., in Fig. 1, treedist(FACT1, "ARG1: her thanks") = 1, treedist(FACT1, "ARG0: people") = 2, treedist(FACT2, "ARG0: people") = 1.

[9] Using the interface from https://github.com/sheffieldnlp/highres.

[10] Note that Corr-F/A are calculated with content weighting with respect to the reference. Therefore, strictly speaking, Corr-F/A are different to all existing metrics but still share some properties with them. We show the correlation between Corr-F/A and existing metrics in terms of relative system ranking, rather than a head-to-head metrics comparison.

[11] For the first two, we use candidate summaries provided by the authors. For the third, we generated summaries by training a model with code and data offered by the authors.

[12] We allow 4 chunks of max. 50 words total for fact-level and 5 chunks of max. 20 words for argument-level annotation.

Granularity	PCC-W	PCC-S
Argument-level	0.3326	0.4762
Fact-level	0.3129	0.7291

Table 1: Correlation of automatic content weighting and selection with human highlights.

Fact Level: By adapting Eq. 5, we calculate a weight w^t_{ki} for each token in document D w.r.t. bolded fact F^S_i in the summary S. The weight w^f_{ij} between fact F^D_i in the document and F^S_j in its summary is calculated using Eq. 6. We use Eq. 3 to get the manual fact content weighting $\mathbf{w}^f_{man}$.

5.2 Agreement with Manual Weighting

Correlation: We evaluate how automatic content weighting $\mathbf{w}^a_{gold}$ and $\mathbf{w}^f_{gold}$ correlates with manual content weighting $\mathbf{w}^a_{man}$ and $\mathbf{w}^f_{man}$. Using the Pearson Correlation Coefficient directly over the content weights (PCC-W), we evaluate the correlation between content weights assigned by human judges and automatically calculated weights – PCC($\mathbf{w}^*_{gold}, \mathbf{w}^*_{man}$). As a more extreme form of weighting, we compute the correlation between content *"selected"* (i.e. ignoring computed weights) by human judges and the automatic mechanism (PCC-S); we set the value to 1 if the weight is over 0, meaning the fact/argument is selected.

While content-weighting correlations are just moderate, content-selection correlations are strong, especially the fact-based (Table 1). In other words, the automatic method attends to facts human judges consider important, but weighs them differently.

System-level Agreement: We check system-level agreement on Corr-F and Corr-A metrics when using automatic vs. manual content weighting (Table 2): We compute fact/argument-level content weights $\mathbf{w}^*_{cand}$ for each system (cf. Section 4). We then calculate Corr-F and Corr-A of $\mathbf{w}^*_{cand}$ against both $\mathbf{w}^*_{man}$ (manual weighting) and $\mathbf{w}^*_{gold}$ (automatic weighting) on the 50 articles with human annotation introduced in Section 5.1.

The Corr-F metric shows the same system-level ordering for both manual and automatic content weighting. Furthermore, both manual and automatic content weighting agree that TCONVS2S and PTGEN achieve similar performance but are strongly outperformed by BERTSUMABS.

5.3 Comparison to existing metrics

Corr-F/A vs. referenceless metrics: HROUGE score (Hardy et al., 2019) is a content-weighting-based referenceless evaluation metric. Unlike our approach, it operates on token level and is entirely based on manual annotation. The evaluation results in Table 3 show that Corr-F/A's ranking is identical to HROUGE's unigram and bigram precision, with Corr-F also assigning similar proportions.[13]

Model	Corr-F	Corr-A
Manual content weighting – $\mathbf{w}^*_{cand}$ vs. $\mathbf{w}^*_{man}$		
TCONVS2S	0.2274	0.2464
PTGEN	0.2180	0.2433
BERTSUMABS	**0.2508**	**0.2662**
Automatic content weighting – $\mathbf{w}^*_{cand}$ vs. $\mathbf{w}^*_{gold}$		
TCONVS2S	0.6203	0.6280
PTGEN	0.5822	0.5727
BERTSUMABS	**0.6714**	**0.6533**

Table 2: System-level scores for manual and automatic content weighting on 50 human-annotated documents.

Model	Unigram		Bigram	
	Pre	Rec	Pre	Rec
TCONVS2S	7.64	5.37	3.16	2.08
PTGEN	7.62	**6.42**	3.25	**2.61**
BERTSUMABS	**8.24**	6.25	**3.29**	2.41

Table 3: HROUGE on 50 human-annotated documents.

Corr-F/A vs. reference-based metrics: ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) are both reference-based metrics, which compute a similarity score for each token in the candidate sentence with each token in the reference sentence. However, instead of exact matches as used in ROUGE, BERTScore computes token similarity using contextual embeddings. Comparing to ROUGE and BERTScore on the full XSum test set (see Table 4) shows full agreement on system ordering for both metrics.

6 Discussion

6.1 Error Analysis

We now provide examples demonstrating the strength and weaknesses of Corr-F/A by analysing system outputs where BERTScore and Corr-F/A demonstrate different ordering.

Strengths: (1) **Corr-F/A are more sensitive to content-level hallucination than BERTScore.** Summaries with facts/arguments never mentioned in the original document get much lower Corr-F/A scores than summaries with content that appears in the document verbatim or as a paraphrase. Example 1 in Table 5 shows Corr-F/A penalizing the incorrect fact "to become the next president" generated by BERTSUMABS, while giving higher scores to TCONVS2S which paraphrased "abdicate" with

[13] We computed HROUGE for BERTSUMABS using https://github.com/sheffieldnlp/highres.

Model	CorrF/A		CorrF/A(L)		ROUGE			BERTScore		
	Corr-F	Corr-A	Corr-F	Corr-A	R1	R2	RL	P	R	F1
TConvS2S	0.616	0.636	0.700	0.650	31.89	11.54	25.75	0.613	0.573	0.591
PtGen	0.596	0.623	0.664	0.620	29.70	9.21	23.24	0.577	0.566	0.570
BertSumAbs	**0.655**	**0.683**	**0.715**	**0.670**	**38.53**	**16.09**	**30.80**	**0.628**	**0.616**	**0.621**

Table 4: Summarisation models evaluated using Corr-F/A on full test set, with ROUGE and BERTScore scores. Note that Corr-F/A(L) is Corr-F/A calculated using a lower-performing SRL tool (He et al., 2017, see Section 6.2).

#	Source	Summary	Corr-F	Corr-A	BS-F1
	Ground truth	Japan's emperor Akihito has expressed his desire to abdicate in the next few years, public broadcaster NHK reports.			
1	BertSumAbs	Japan's emperor Akihito is considering whether to become the next president of the country, reports say.	0.68	0.68	0.67
	TConvS2S	Japan's emperor Akihito has announced that he will step down in the Japanese capital, Tokyo.	0.81	0.71	0.67
	Ground truth	Dick Advocaat has resigned as Sunderland boss, with the team yet to win in the Premier League this season.			
2	BertSumAbs	Sunderland manager Dick Advocaat has left the club by mutual consent after only eight games in charge.	0.60	0.66	0.65
	PtGen	Sunderland have appointed former boss Dick Advocaat as manager at the end of the season to sign a new deal.	0.26	0.34	0.65
	Ground truth	A Chinese space capsule carrying three crew members has returned to Earth following a 13-day mission.			
3	BertSumAbs	China has successfully landed its first ever space flight, in a move hailed as a "historic moment".	0.56	0.67	0.53
	TConvS2S	China has successfully launched the first ever robotic mission to date for the first time in its history.	0.85	0.68	0.51
	Ground truth	A council plans to employ its own staff to help young people with mental health problems.			
4	BertSumAbs	A new academy to train people with mental health problems is to be set up in West Berkshire.	0.82	0.68	0.64
	TConvS2S	A new academy for children with mental health problems is being launched in West Berkshire.	0.73	0.56	0.67

Table 5: Examples of system outputs where Corr-F/A and BERTScore-F1 disagree on system ordering.

"step down". (2) **Corr-F/A better identify paraphrases**, especially those containing extra content mentioned in the document but not in the ground-truth summary. Example 2 in Table 5 shows that Corr-F/A do not penalize BertSumAbs for generating the argument "after only eight games in charge", which is mentioned in the document.

Weaknesses: (1) **Corr-F is weaker in identifying token-level hallucination,**[14] as in Example 3 in Table 5. Corr-F gives a higher score to TConvS2S output with one hallucinated token "robotic". However, Corr-A's more fine-grained approach works slightly better in this case. (2) **Corr-F/A tend to under-score summaries containing content mentioned in the ground truth but only touched briefly in the document.** In Example 4 in Table 5, Corr-F/A score the output of TConvS2S lower, even though it correctly captures "an academy for children with mental health", which is mentioned only once in the document.

In sum, Corr-F/A is less dependent on the reference summary by also considering the source document, and thus has less of a reference bias than BERTScore. In addition, Corr-F/A helps to identify ungrounded facts, i.e. content-level hallucinations, which is important for identifying misinformation in automated news reporting.

6.2 Robustness of Corr-F/A

As noted in Section 3.1, Corr-F/A is based on publicly available SRL tools. To demonstrate the robustness of our metrics, we evaluate the same system outputs with Corr-F/A calculated using a lower-performing SRL tool (He et al., 2017).[15] The results are shown as Corr-F/A(L) in Table 4 and show full agreement with Corr-F/A in terms of system ordering. However, the better performing original SRL system widens the margin between systems.

7 Conclusions and Future Work

We present an automatic evaluation framework for abstractive summarisation, which is low-cost and robust, as it does not rely on expert annotators nor is susceptible to crowdsourcing noise. Using fact representations, we are able to capture semantically similar, but at the same time distant in surface form, content in the summary that aligns with arbitrarily far-apart parts of the input document, casting our metric to be directly interpretable. Our metric is more sensitive to perturbations of the facts in the target summary, which resemble common hallucination phenomena of neural decoders (see Figure 2-3 in Appendix A for examples). In the future, we intend to investigate different meaning representation formalisms, such as AMR (Banarescu et al., 2013) and Dynamic Syntax (Kempson et al., 2001) and extend to other datasets (e.g. multiple-reference summarization) and tasks (e.g. response generation in dialogue).

Acknowledgements

This research received funding from the EPSRC project MaDrIgAL (EP/N017536/1) and Charles University project PRIMUS/19/SCI/10. We would like to acknowledge the AWS Cloud Credits for Research programme.

[14] Token-level hallucination means an incorrect token within an otherwise correct fact structure. Content-level hallucination happens when whole facts or arguments are hallucinated.

[15] F1 on the Ontonotes 5.0 dataset is 81.6% ($\delta = -4.89$).

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the Original: Fact Aware Neural Abstractive Summarization. In *AAAI*, New Orleans, LA, USA. ArXiv: 1711.04434.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In *ACL*, Florence, Italy.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123 – 156.

Ben Goodrich, Vinay Rao, Mohammad Saleh, and Peter J. Liu. 2019. Assessing The Factual Accuracy of Generated Text. In *KDD*, Anchorage, AK, USA. ArXiv: 1905.13322.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369, Melbourne, Australia.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Ruth M Kempson, Wilfried Meyer-Viol, and Dov M Gabbay. 2001. *Dynamic syntax: The flow of language understanding*. Blackwell Oxford.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3728–3738, Hong Kong, China.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium.

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*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium.

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: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers

Unite! Unsupervised Metrics for Reinforced Summarization Models. In *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) and 9th International Joint Conference on Natural Language Processing (IJCNLP)*, Hong Kong. ArXiv: 1909.01610.

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*, pages 1073–1083, Vancouver, Canada.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing Lightweight Pyramids for Manual Summary Evaluation. In *NAACL*, Minneapolis, MN, USA. ArXiv: 1904.05929.

Simeng Sun and Ani Nenkova. 2019. The Feasibility of Embedding Based Automatic Evaluation for Single Document Summarization. In *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) and 9th International Joint Conference on Natural Language Processing (IJCNLP)*, Hong Kong.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 419–426, Barcelona, Spain. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. SUM-QE: a BERT-based Summary Quality Estimation Model. 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 6007–6013, Hong Kong, China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

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 *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) and 9th International Joint Conference on Natural Language Processing (IJCNLP)*, Hong Kong. ArXiv: 1909.02622.

A Fact-level Content Weighting Examples

Fig. 2 and 3 show examples for documents weighted using Corr-F/Corr-A with respect to different summaries.

In Fig. 2, the left column shows one document weighted by the reference summary and two system-generated summaries from BERTSUMABS and TCONVS2S respectively (summaries are shown in the right column). As we can see, there are 4 relatively important facts in the document weighted by the reference summary. BERTSUMABS and TCONVS2S capture 3 and 2 out of 4, respectively. Other than the important facts highlighted by the reference summary, TCONVS2S also assigns high weights to other facts; that leads to the hallucinated generation and lower Corr-F Corr-A scores. On the other hand, BERTSUMABS's summary weighs facts in the document in a similar way to the reference summary, which lead to a strongly related summary and high Corr-F and Corr-A scores.

In Fig. 3, there are 5 relatively important facts in the document weighted by the reference summary. BERTSUMABS and TCONVS2S capture 4 and 3 out of 5, respectively. Both systems miss the fact "Pope Francis, who has taken a more liberal stance on homosexuality". However, the weight of this fact given by BERTSUMABS's output is higher than with TCONVS2S's. The Corr-F and Corr-A are lower for TConvS2S due to misweighting of informative facts in the document.

B Annotation Interface

We provide the following illustrations of the human annotation interface:

- Annotation interface for manual content weighting examples, including the instructions, for fact-level (Fig. 4 and 5) and argument-level (Fig. 6 and 7) annotation,
- Examples of human annotation results for fact (Fig. 9) and argument (Fig. 8) level.

Please refer to the individual figure captions for detailed descriptions.

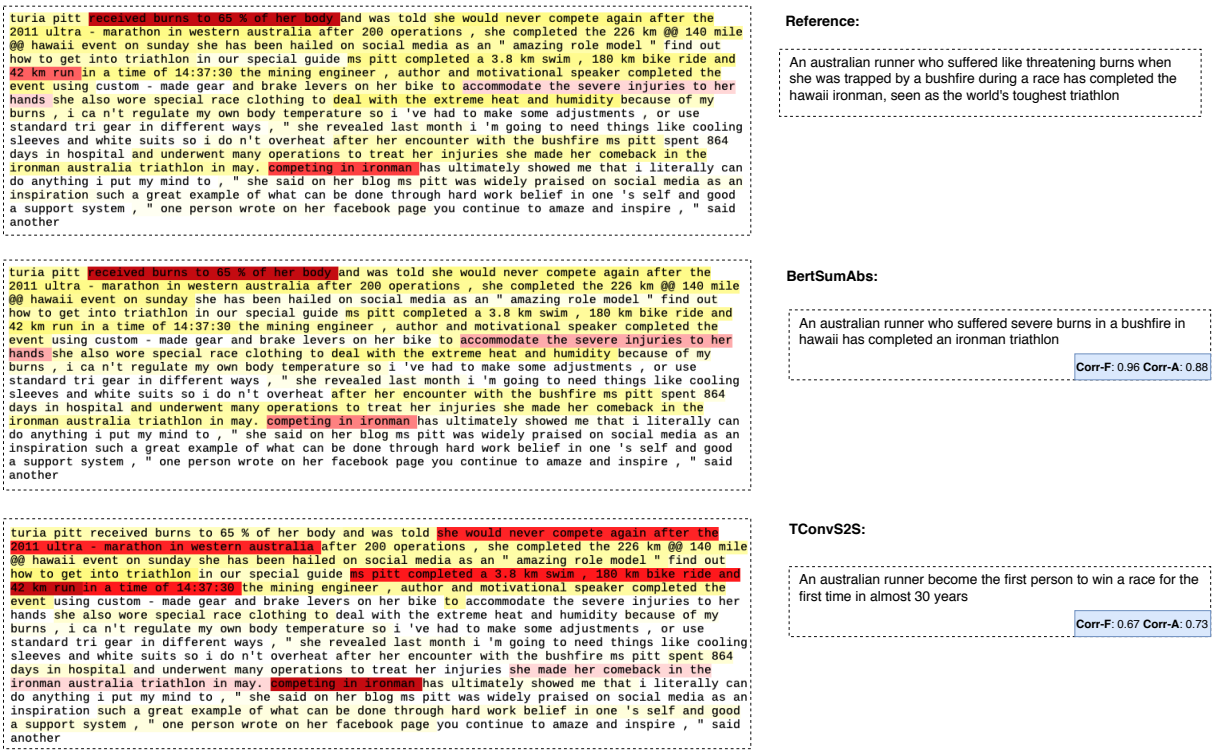

Figure 2: A document (left) weighted with respect to a reference summary and two system outputs (right), with Corr-F/Corr-A scores. The colour represents the sum of argument- and fact-level weights for each token (Eqs. 3 and 4). The darker the colour, the more important the fact is.

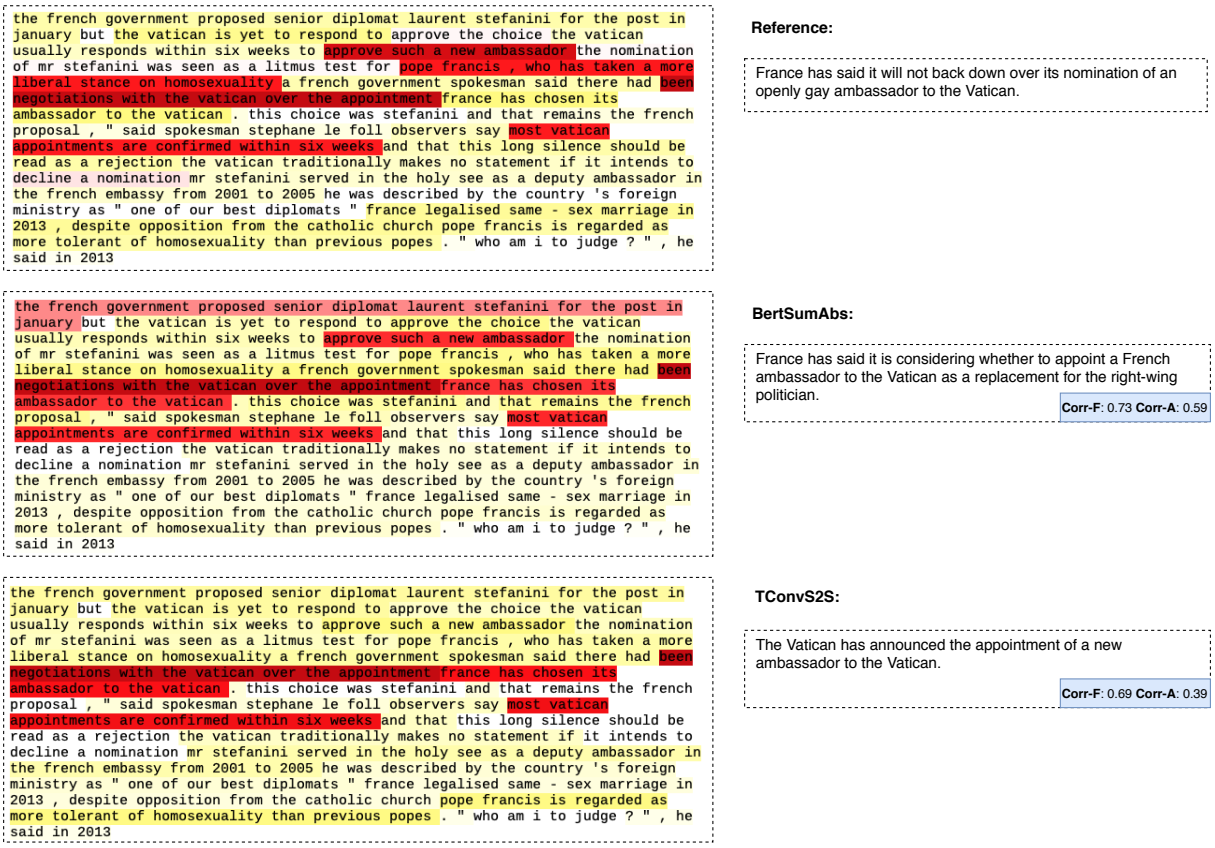

Figure 3: Another document (left) weighted with respect to a reference summary and two system outputs (right), with Corr-F/Corr-A scores (see Fig. 2 for details).

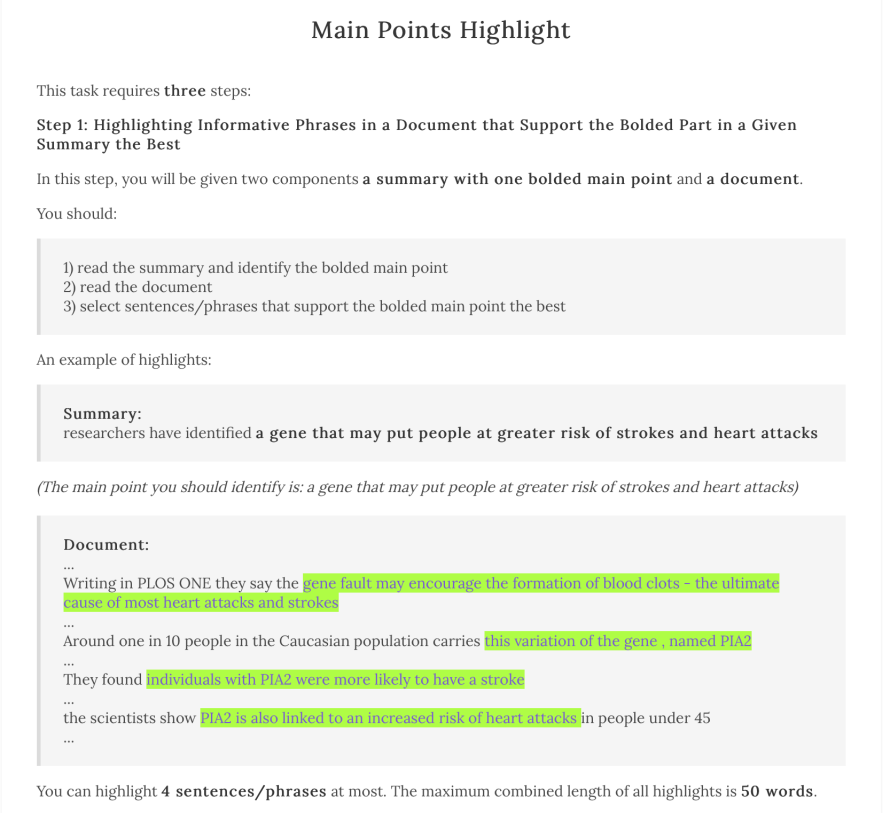

Figure 4: The instruction for fact-level human highlight annotation.

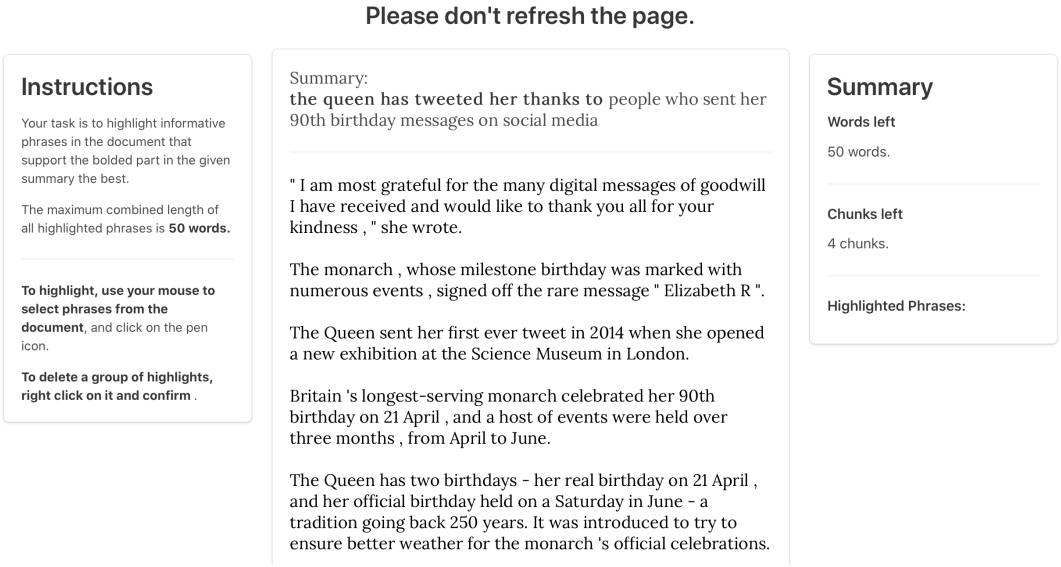

Figure 5: The human annotation interface for fact level. Human judges are required to highlight content in the document that is supporting the fact printed in bold "The Queen has tweeted her thanks" (FACT1 of the summary in Figure 1 in the paper).

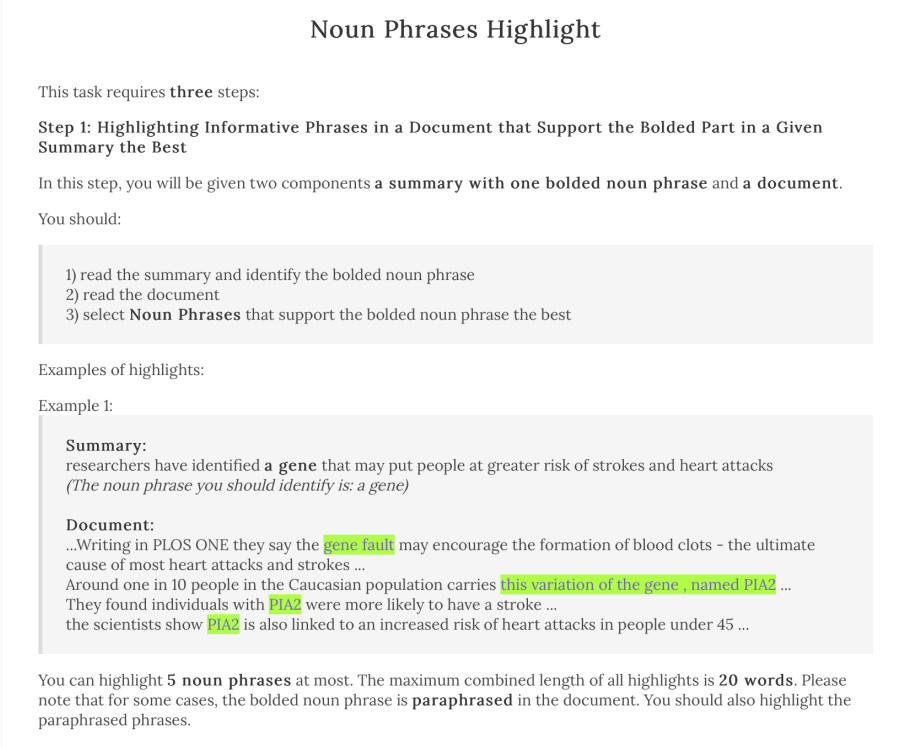

Figure 6: The instruction for argument-level human highlight annotation.

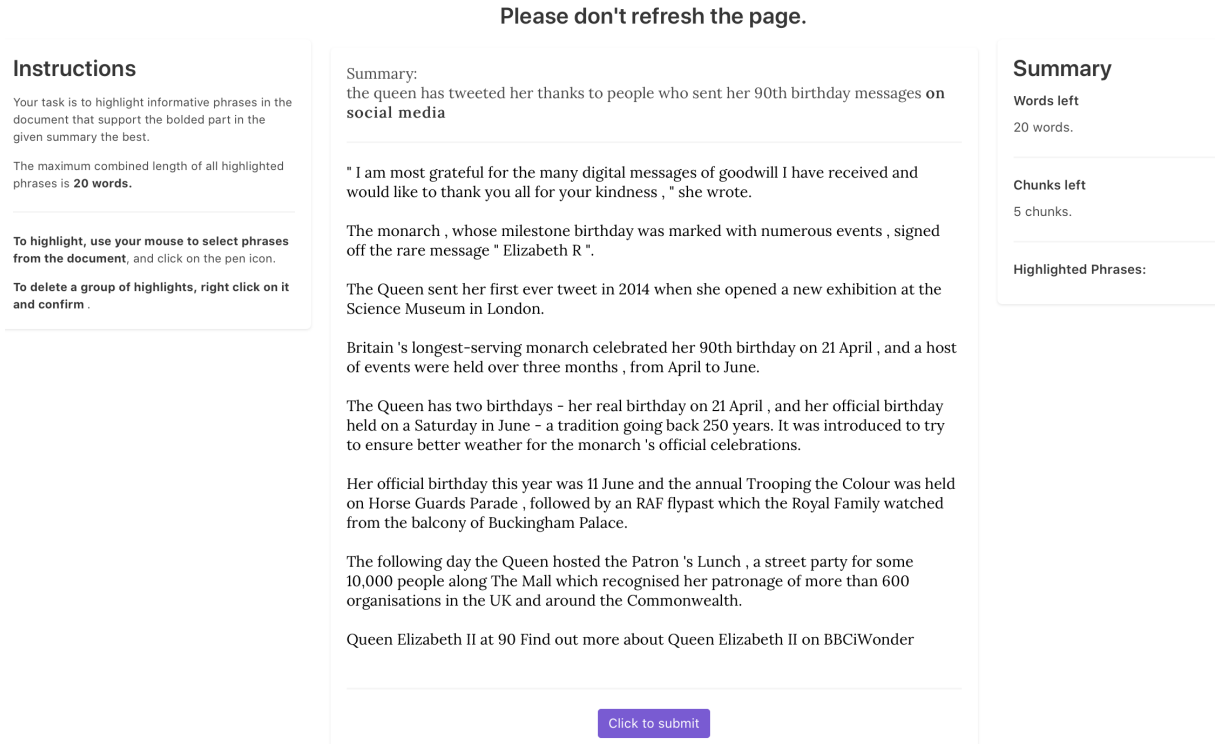

Figure 7: The human annotation interface for argument level. Human judges are required to highlight content in the document that is supporting the phrase printed in bold “on social media” (argument ARGM-LOC of FACT2 of the summary in Figure 1 in the paper).

Article

" i am most grateful for the many digital messages of goodwill i have received and would like to thank you all for your kindness , " she wrote . the monarch , whose milestone birthday was marked with numerous events , signed off the rare message " elizabeth r " . the queen sent her first ever tweet in 2014 when she opened a new exhibition at the science museum in london . britain 's longest-serving monarch celebrated her 90th birthday on 21 april , and a host of events were held over three months , from april to june . the queen has two birthdays - her real birthday on 21 april , and her official birthday held on a saturday in june - a tradition going back 250 years . it was introduced to try to ensure better weather for the monarch 's official celebrations . her official birthday this year was 11 june and the annual trooping the colour was held on horse guards parade , followed by an raf flypast which the royal family watched from the balcony of buckingham palace . the following day the queen hosted the patron 's lunch , a street party for some 10,000 people along the mall which recognised her patronage of more than 600 organisations in the uk and around the commonwealth . queen elizabeth ii at 90 find out more about queen elizabeth ii on bbc iwonder

Figure 8: Human highlight annotation for the argument ARG1 of FACT1 “her thanks” of the summary in Figure 1 in the paper.

Article

" i am most grateful for the many digital messages of goodwill i have received and would like to thank you all for your kindness , " she wrote . the monarch , whose milestone birthday was marked with numerous events , signed off the rare message " elizabeth r " . the queen sent her first ever tweet in 2014 when she opened a new exhibition at the science museum in london . britain 's longest-serving monarch celebrated her 90th birthday on 21 april , and a host of events were held over three months , from april to june . the queen has two birthdays - her real birthday on 21 april , and her official birthday held on a saturday in june - a tradition going back 250 years . it was introduced to try to ensure better weather for the monarch 's official celebrations . her official birthday this year was 11 june and the annual trooping the colour was held on horse guards parade , followed by an raf flypast which the royal family watched from the balcony of buckingham palace . the following day the queen hosted the patron 's lunch , a street party for some 10,000 people along the mall which recognised her patronage of more than 600 organisations in the uk and around the commonwealth . queen elizabeth ii at 90 find out more about queen elizabeth ii on bbc iwonder

Figure 9: Human highlight annotation for the FACT1 “The Queen has tweeted her thanks” of the summary in Figure 1 in the paper.