Event-based evaluation of abstractive news summarization

Huiling You[1], **Samia Touileb**[2], **Erik Velldal**[1], **and Lilja Øvrelid**[1]
[1]University of Oslo
[2]University of Bergen
{huiliny, erikve, liljao}@ifi.uio.no
samia.touileb@uib.no

Abstract

An abstractive summary of a news article contains its most important information in a condensed version. The evaluation of automatically generated summaries by generative language models relies heavily on human-authored summaries as gold references, by calculating overlapping units or similarity scores. News articles report events, and ideally so should the summaries. In this work, we propose to evaluate the quality of abstractive summaries by calculating overlapping events between generated summaries, reference summaries, and the original news articles. We experiment on a richly annotated Norwegian dataset comprising both events annotations and summaries authored by expert human annotators. Our approach provides more insight into the event information contained in the summaries.

1 Introduction

A summary of a news article provides a condensed version of its main content (El-Kassas et al., 2021). One of the primary practical applications of large language models (LLMs) is generating concise text summaries, and many news publishers in Norway have already integrated LLM-generated summaries into their articles. However, assessing the quality and accuracy of these summaries remains a challenge. Current evaluation metrics compare generated summaries to ideal summaries created by humans, in terms of overlapping words/units, such as ROUGE-L (Lin, 2004), or semantic similarity, such as BERTScore (Zhang* et al., 2020). However, these metrics provide limited information on the semantic content of the summaries themselves.

With the increasing usage of LLMs for text generation, there has been a growing number of studies on evaluating the factuality of these texts from the perspective of contained information, such as FACTSCORE (Min et al., 2023). For summarization, Zhang and Bansal (2021) propose to use semantic triplet units as a judgment of the semantic content units in generated texts, and Liu et al. (2023) also propose a similar protocol based on semantic units, named Atomic Content Units. Inspired by event extraction (EE), a NLP task that extracts event information from unstructured texts into structured forms (Doddington et al., 2004), we propose to analyze the quality of news article summaries by comparing the overlapping events between generated summaries, reference summaries, and the source articles. By using structured event information, we provide more insight into both the generated summaries and human-authored summaries. We experiment on a Norwegian dataset with rich annotations both for events (EDEN (Touileb et al., 2024)), and summaries (NorSumm (Touileb et al., 2025)), and demonstrate the usefulness of the proposed event-based evaluation metric which is grounded in the overlap of identified events.

2 Event-overlap

Our proposed metric calculates the degree of overlapping events between summaries (generated and human-authored) and the source texts. First, an event extraction system is used to extract events from summaries and source articles. Second, standard event extraction evaluation metrics are adapted and applied to calculate the actual event overlaps.

2.1 Event extraction

An event (Doddington et al., 2004) contains four key elements: 1) **event type** is the specific type of event defined within an ontology; 2) **event trigger** is the word(s) in the text that describes the event; 3) **event argument** is the attribute and actual participant of an event in the text; 4) **argument role** is the role played by an argument in the specific event. Figure 1 shows an example of a Norwegian

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 504–510
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

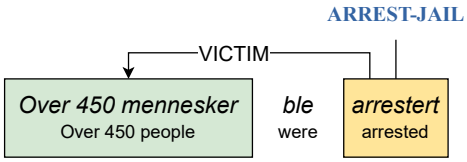

Figure 1: Example of a sentence with event annotation. The ARREST-JAIL event has the trigger "arrested", and the VICTIM argument is "Over 450 people".

sentence annotated for an ARREST-JAIL event with "arrested" as the event trigger, and a VICTIM argument "Over 450 people". We use an existing event extraction system NorEventGen (You et al., 2025) to obtain event information in these structured formats.

We perform event extraction on three different texts: 1) model-generated summaries; 2) human-authored summaries; and 3) original news articles.

2.2 Event-overlap analysis

Our event-overlap metric is adapted from the classical evaluation metrics of event extraction (Lin et al., 2020; Nguyen et al., 2021), as follows: an event trigger is correctly identified (Trg-I) if its offsets match a reference trigger, and correctly classified (Trg-C) if its event type also matches a reference trigger; An argument is correctly identified (Arg-I) if its offsets match a reference argument, and correctly classified (Arg-C) if its argument role also matches the reference argument.

Since an abstractive summary does not perform text extraction from the source article, we do not expect a perfect match between an event trigger / argument from the summary and one from the article. As an alternative, we use BERTScore (Zhang* et al., 2020) as a reference to check if two pieces of texts are similar.[1] Unlike in event extraction, we prioritize the labels, namely event type and argument role. We do not take trigger word(s) into account, because the event type information itself is sufficient, and unlike event arguments, which are named entities, trigger words are more often rephrased with a different choice of words in summaries. With the corresponding adaptation, our proposed event-overlap metric calculates the following three categories of scores:

- An event type (eType-C) overlaps if it exists in both lists of extracted events.
- An argument role (Role-C) overlaps if the event type and argument role overlap.
- An argument (Arg-C) overlaps if the event type, argument role, and argument word(s) overlap.

The Precision (P), Recall (R), and F1 scores of each category are calculated. The final event-overlap score is an aggregated score of the three categories of scores: **Event-overlap** = **Average**([eType-C, Role-C, Arg-C]). Depending on the event overlap of different texts, different scores are used:

- **Event-overlap between summaries**: the final event-overlap score is the average Recall scores of eType-C, Role-C, and Arg-C. Recall scores prioritize the events that are in the gold summaries.
- **Event-overlap between summaries and original articles**: the final event-overlap score is the average Precision scores of eType-C, Role-C, and Arg-C. Precision scores provide evaluation of identified events in the summaries that are also present in the original articles.

3 Experimental setup

Datasets We use two recently released datasets: the Norwegian event detection dataset EDEN (Touileb et al., 2024) and the human-authored summaries of Norwegian news articles dataset NorSumm (Touileb et al., 2025). The source articles of NorSumm are a subset of EDEN. These parallel annotations of events and summaries make it possible to evaluate our approach and contrast gold vs predicted event information on gold vs generated summaries. More concretely, we use the test set of NorSumm, which contains 33 news articles, each coupled with three unique human-authored summaries.

LLMs For automatic summarization, we evaluate a range of Norwegian and Nordic open-source pretrained and instruction-finetuned decoder-only LLMs: Llama-3-8B-instruct,[2] Llama-3-8B,[3] Meta-Llama-3-8B-Instruct[4], Mistral-

[1] We use a heuristic threshold of 0.7. If the BERTScore is larger than 0.7, two text snippets will be considered similar, the same as perfect match in event extraction metric.

[2] https://huggingface.co/AI-Sweden-Models/Llama-3-8B-instruct

[3] https://huggingface.co/AI-Sweden-Models/Llama-3-8B

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Nemo-Instruct-2407,[5] Normistral-11b-warm[6], and Normistral-7b-warm-instruct.[7] All the LLMs are available via HuggingFace.[8] We use the same prompts as in the NorSumm evaluation (Touileb et al., 2025) to generate summaries, and keep only one summary that has highest average score of ROUGE-L and BERTScore values for each model.

Event extraction system We use a generative event extraction system NorEventGen (You et al., 2025) to identify and extract events from both the original articles and the summaries. NorEventGen is trained on EDEN, and holds the current SOTA results. The system performs sentence-level extraction. In our experiments, both the original articles and the summaries are first split into sentences, and then event prediction is performed on each of the sentences.

4 Results and discussion

We here present the analysis of our event-overlap metric on the test set of NorSumm. We first present the event-overlap between summaries and the original articles; we then present the event-overlap between generated summaries and human-authored summaries. Finally, we discuss the overall picture summarizing event-overlap scores.

4.1 Event-overlap between summaries and the original articles

Table 1 shows the event-overlap between the summaries (both human-authored and generated) and the original articles. As the results show, both generated summaries and human-authored summaries generally discuss events that are described in the original articles, and there are always fewer events in the summaries. As the Precision scores of eType-C are always above 90%, it is rare for events that are not discussed in the source article to be mentioned in the summary, which is especially true for generated summaries. The Recall scores of eType-C are much lower, meaning that there are far fewer events in the summaries; the number of events varies considerably among generated summaries. The Precision scores of Role-C and Arg-C show that events are discussed with different levels of detail in the summaries compared to the news articles. Similarly, the event-overlap metric shows that Normistral-11b-warm is the best-performing model, but the summaries generated by Llama-3-8B and Normistral-7b-warm-instruct also produce relatively good results with each of the fine-grained metrics.

Table 3 provides detailed event statistics of both human-authored and generated summaries, together with event information of the original articles. In general, there are always fewer events in the summaries as compared to in the original articles, which is expected. Human annotators have rather high agreement on event numbers, but the number of argument roles vary quite a lot, meaning they tend to describe the events with varied details when writing the summaries. For model-generated summaries, some describe considerably more events than others; the summaries generated by Normistral-7b-warm-instruct contain twice the number of events compared with the summaries generated by Llama-3-8B-instruct.

Instead of predicted events, we can also assess the influence of event detection accuracy and compare the gold event annotation of the original articles to calculate the event-overlap scores. As Table 2 shows, the event-overlap scores are still relatively high, similar to using predicted events of the articles. The drops in scores are expected, because the event extraction model is not perfect and less frequent events are annotated, which would normally not be included in the summary.

With gold events, the ranking of the models turns out to be different from when predicted events are used; summaries generated by Meta-Llama-3-8B-Instruct have the highest event-overlap score with the original articles, instead of Normistral-11b-warm. However, the top-performing models remain quite similar.

4.2 Event-overlap between summaries

Table 4 shows the event-overlap between model-generated summaries and human-authored summaries. As the event-overlap scores show, the proportion of shared events in generated summaries with reference summaries varies across the various models. In general, eType-C scores are much higher than Role-C and Arg-C scores, indicating that the same events are discussed with different details. Table 5 presents an example of a `TRANSFER-OWNERSHIP` event described in a human-authored summary and a model-generated sum-

[5] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

[6] https://huggingface.co/norallm/normistral-11b-warm

[7] https://huggingface.co/norallm/normistral-7b-warm-instruct

[8] https://huggingface.co/models

Summary	eType-C			Role-C			Arg-C			Event-overlap
	P	R	F1	P	R	F1	P	R	F1	
Human-authored	90.7	13.4	23.4	84.7	13.2	22.8	68.2	10.7	18.4	81.2
Llama-3-8B-instruct	93.3	8.3	15.3	87.3	6.8	12.6	70.4	5.5	10.2	83.7 $_{(6)}$
Llama-3-8B	**98.4**	12.1	21.5	89.2	10.8	19.3	81.1	9.9	17.6	89.6 $_{(2)}$
Meta-Llama-3-8B-Instruct	97.8	8.9	16.3	90.0	7.9	14.5	76.3	6.7	12.3	88.0 $_{(3)}$
Mistral-Nemo-Instruct-2407	98.0	9.5	17.3	87.1	8.1	14.8	69.4	6.5	11.8	84.8 $_{(4)}$
Normistral-11b-warm	96.7	**17.2**	**29.2**	**90.8**	16.2	27.5	**82.2**	**14.7**	**24.9**	**89.9** $_{(1)}$
Normistral-7b-warm-instruct	94.6	**17.2**	29.1	88.5	**16.9**	**28.3**	69.5	13.3	22.3	84.2 $_{(5)}$

Table 1: Event-overlap between summaries and the original articles, with event prediction is performed with NorEventGen. The subscripts indicate the corresponding ranking of the model based on the score.

Summary	eType-C			Role-C			Arg-C			Event-overlap
	P	R	F1	P	R	F1	P	R	F1	
Human-authored	74.2	13.1	22.4	69.4	11.9	20.4	59.2	10.2	17.4	67.6
Llama-3-8B-instruct	84.4	9.0	16.2	76.1	6.5	12.0	66.2	5.7	10.5	75.6 $_{(4)}$
Llama-3-8B	82.3	12.1	21.0	76.6	10.3	18.1	68.5	9.2	16.2	75.8 $_{(3)}$
Meta-Llama-3-8B-Instruct	**87.0**	9.5	17.1	82.5	8.0	14.6	**75.0**	7.3	13.3	**81.5** $_{(1)}$
Mistral-Nemo-Instruct-2407	83.7	9.7	17.4	**83.5**	8.6	15.6	74.1	7.6	13.8	80.4 $_{(2)}$
Normistral-11b-warm	80.0	17.0	28.1	77.3	15.3	25.5	69.3	**13.7**	**22.9**	75.5 $_{(5)}$
Normistral-7b-warm-instruct	87.0	**18.9**	**31.1**	77.0	**16.2**	**26.8**	59.8	12.6	20.8	74.6 $_{(6)}$

Table 2: Event-overlap between summaries (predicted events) and the original articles (gold events). The subscripts indicate the corresponding ranking of the model based on the score.

Summary	#Events	#Roles	#Event types	#Role types
Annotator_1	77	156	17	23
Annotator_2	77	146	16	20
Annotator_3	71	126	16	24
Llama-3-8B-instruct	45	71	13	17
Llama-3-8B	62	111	14	19
Meta-Llama-3-8B-Instruct	46	80	14	20
Mistral-Nemo-Instruct-2407	49	85	12	19
Normistral-11b-warm	90	163	15	20
Normistral-7b-warm-instruct	92	174	15	23
Gold events in original articles	423	826	23	25
Predicted events in original articles	506	918	23	25

Table 3: Event statistics of human-authored summaries by three different annotators and generated summaries by different models. Events are predicted with the selected event extraction system.

Model	ROUGE-L	BERTScore	eType-C			Role-C			Arg-C			Event-overlap
			P	R	F1	P	R	F1	P	R	F1	
Llama-3-8B-instruct	24.5 $_{(6)}$	72.1 $_{(6)}$	74.1	44.6	55.7	58.2	29.4	39.0	45.1	22.9	30.3	32.3 $_{(6)}$
Llama-3-8B	36.7 $_{(3)}$	73.3 $_{(4)}$	74.7	61.9	67.7	61.3	48.0	53.7	44.7	35.0	39.2	48.3 $_{(3)}$
Meta-Llama-3-8B-Instruct	28.8 $_{(5)}$	75.2 $_{(2)}$	**75.4**	46.3	57.3	**62.5**	35.3	45.0	**52.9**	29.8	38.1	37.1 $_{(4)}$
Mistral-Nemo-Instruct-2407	**41.1** $_{(1)}$	**75.8** $_{(1)}$	67.4	43.9	53.2	55.7	33.0	41.4	45.5	27.0	33.8	34.6 $_{(5)}$
Normistral-11b-warm	34.9 $_{(4)}$	73.1 $_{(5)}$	70.4	**84.6**	**76.8**	55.6	**63.9**	**59.4**	40.3	46.1	**42.9**	**64.9** $_{(1)}$
Normistral-7b-warm-instruct	37.8 $_{(2)}$	73.7 $_{(3)}$	64.5	79.2	71.1	51.0	62.6	56.1	37.9	**46.5**	41.7	62.8 $_{(2)}$

Table 4: Event-overlap between generated summaries and human-authored summaries. The subscripts indicate the corresponding ranking of the model based on the score.

Human-authored	Tommy Sharif sikret seg "Diamanten", toppen av det historiske Holmenkollen-tårnet, før nettauksjonen ble avsluttet kl 16.30 på søndag. *Tommy Sharif secured the "Diamond", the top of the historic Holmenkollen Tower, before the online auction ended at 4:30 p.m. on Sunday.*
Generated	Tommy Sharif sikret seg vinnerbudet på «Diamanten» på Holmenkollen-tårnet da nettauksjonen ble avsluttet søndag. *Tommy Sharif secured the winning bid for the "Diamond" on the Holmenkollen Tower when the online auction ended on Sunday.*

Table 5: Example sentence describing the same event, taken from a human-authored summary and a summary generated by Normistral-11b-warm.

Human-authored	ARREST-JAIL, ATTACK, BE-BORN, CONVICT, DEMONSTRATE, DIE, ELECT, END-ORG END-POSITION, INJURE, MEET, PHONE-WRITE, START-ORG, START-POSITION TRANSFER-MONEY, TRANSFER-OWNERSHIP, TRANSPORT, TRIAL-HEARING
Generated	ARREST-JAIL, ATTACK, BE-BORN, CHARGE-INDICT, CONVICT, DEMONSTRATE, DIE, ELECT END-ORG, END-POSITION, EXECUTE, FINE, INJURE, MEET, PHONE-WRITE, START-ORG START-POSITION, TRANSFER-MONEY, TRANSFER-OWNERSHIP, TRANSPORT, TRIAL-HEARING

Table 6: Event types in human-authored summaries and generated summaries.

mary; the human annotator provides more detail about the ARTIFACT, of which the ownership is transferred, and the TIME of the event, but the model stresses that the BUYER gets a winning bid in the auction.

In terms of event types, there are much fewer event types in the summaries. The event ontology of EDEN defines 34 event types, but only half of the event types exist in the reference summaries and even fewer in generated summaries. As such, only certain event types are often considered as main event types, which are then described in the summary. Table 6 lists all the event types that are described in all human-authored summaries and generated summaries, corresponding to 21 and 18 event types.

Compared to ROUGE-L and BERTScore, the standard summarization evaluation metrics, our event-overlap scores result in slightly different rankings of model performances. According to ROUGE-L and BERTScore, the best-performing model is Mistral-Nemo-Instruct-2407, but our event-overlap metric would identify Normistral-11b-warm as the best-performing model.

4.3 Event-overlap: a combined picture

By analyzing the event-overlap scores between model-generated summaries and their corresponding human-authored counterparts, alongside the event-overlap scores between both types of summaries and the original articles, we can gain deeper insight into how each summarization approach captures the core content of the articles. These event-overlap scores, as presented in Table 4 and 1, reveal a notable trend: summaries generated by LLMs often focus on different events within the article compared to those emphasized by human writers. This pattern holds consistently across all the LLMs evaluated in the study. LLMs and human summarizers tend to have different judgments on what constitutes the main events or key points in a news article, showing that LLMs struggle to accurately identify and convey the main story in complex, real-world texts like news articles.

5 Conclusion

In this article, we introduce a new approach for evaluating abstractive summaries using event identification information. Our proposed event-overlap metric quantifies shared events between generated summaries, human-authored summaries, and the original news articles, offering more insight into the event information of the summaries. In conjunction with standard summarization evaluation metrics, our event-overlap metric adds a valuable dimension to assessing the quality of LLM generated summaries. Experiments conducted on NorSumm, a richly annotated Norwegian dataset, demonstrate the effectiveness and practicality of our method. Our approach is also easily adaptable to other datasets and languages.

Limitations

Our work has the following limitations: 1) we only experiment on a small Norwegian dataset, and the event annotation is on a sentence level, but a summary is a condensed version of the entire article; 2) the selected set of generative LLMs is limited; 3) we make a considerable change to the perfect match of argument words in the original event extraction evaluation metric, and our new equivalent using BERTScore with a heuristic value of 0.7 as threshold, needs further experiments; 4) our event-overlap metric is limited by the event extraction system used, and current event extraction systems are far from being perfect.

Acknowledgments

This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the centers for Research-based Innovation scheme, project number 309339.

References

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

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.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. 2021. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 27–38, Online. Association for Computational Linguistics.

Samia Touileb, Vladislav Mikhailov, Marie Kroka, Lilja Øvrelid, and Erik Velldal. 2025. Benchmarking abstractive summarisation: A dataset of human-authored summaries of norwegian news articles. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies(NoDaLiDa/Baltic-HLT 2025)*, pages 729–738, Tallinn, Estonia.

Samia Touileb, Jeanett Murstad, Petter Mæhlum, Lubos Steskal, Lilja Charlotte Storset, Huiling You, and Lilja Øvrelid. 2024. EDEN: A dataset for event detection in Norwegian news. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5495–5506, Torino, Italia. ELRA and ICCL.

Huiling You, Samia Touileb, Erik Velldal, and Lilja Øvrelid. 2025. Noreventgen: generative event extraction from norwegian news. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies(NoDaLiDa/Baltic-HLT 2025)*, pages 801–811, Tallinn, Estonia.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

A Summary statistics

Writing summaries of news articles is a subjective task. Human annotators can write different sum-

Summary	#Summ.	#Tokens	#Avg.
Annotator_1	33	8,679	263
Annotator_2	33	4,256	129
Annotator_3	33	2,732	83
Llama-3-8B-instruct	33	3,308	100
Llama-3-8B	33	4,331	131
Meta-Llama-3-8B-Instruct	33	3,523	106
Mistral-Nemo-Instruct-2407	33	3,019	91
Normistral-11b-warm	33	6,030	182
Normistral-7b-warm-instruct	33	5,653	171

Table 7: Statistics of human-authored summaries and generated summaries for the test set of NorSumm. "#Summ.": number of summaries; "#Tokens": total number of tokens; "#Avg.": average number of tokens per summary.

maries for the same article. In NorSumm, each article is accompanied with three unique summaries written different annotators, who write in very different styles. As shown in Table 7, Annotator_1 creates the longest summaries, while Annotator_3 creates the shortest summaries. The LLMs also generate varied summaries. As shown in Table 7, some models generate rather short summaries, while some models generate rather long summaries.