Extract, Select and Rewrite: A Modular Sentence Summarization Method

Shuo Guan* [1], Vishakh Padmakumar* [2]

[1]UBS AG, [2]New York University

Motivation

- While end-to-end models are dominating text generation tasks today, modular or pipelined approaches have the advantage of greater controllability and better low-resource performance.
- The "modularity" in existing methods rely on extra encoder, as opposed to splitting the model into different modules explicitly, which may limit their interpretability.

We present **ESR**, a three-phase extract-select-rewrite framework for abstractive sentence summarization.

Framework

We decompose summarization into three stages, with knowledge triples as the granularity. The structure of ESR is shown in Figure 1 using a sentence example from Gigaword dataset.

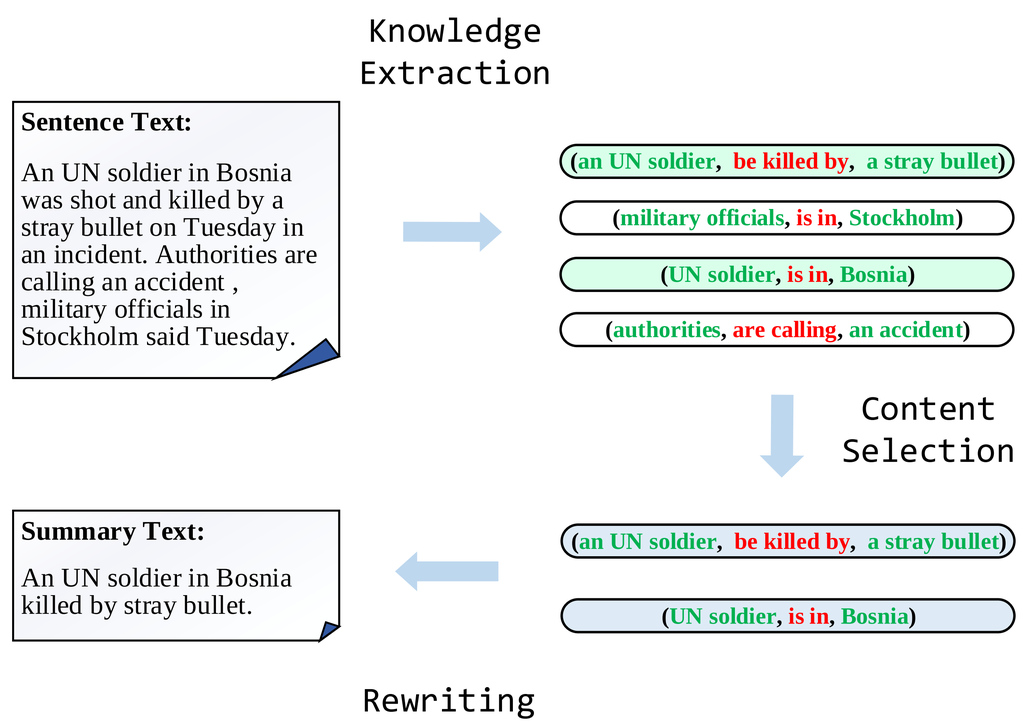
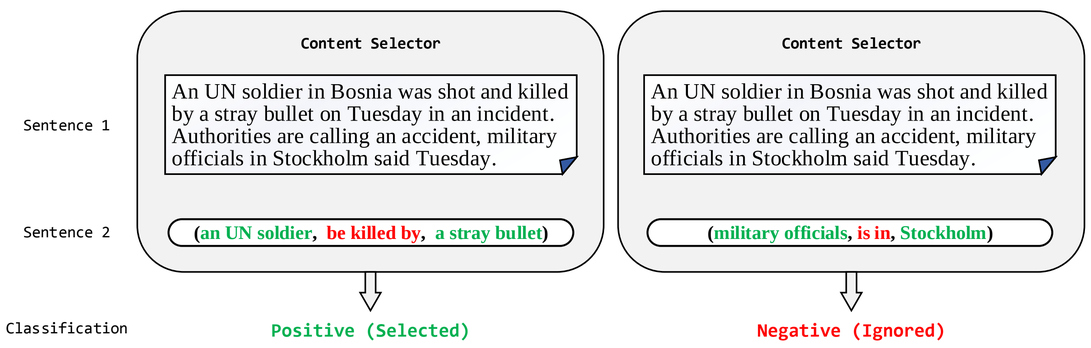
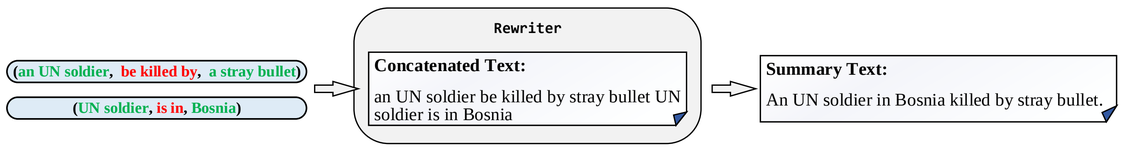


Figure 1. The structure of ESR with the three phases marked.

- **Knowledge Extraction**, where we extract relation triples from the text using off-the-shelf tools. The knowledge triples are in the form of `<entity 1, relation, entity 2>`. We perform extra steps for deleting the overlapping triples.
- **Content Selection**, where a subset of triples are selected. We train it as a sentence-pair classifier with two inputs, the *document* and the *candidate knowledge triple* extracted from it, and an output of whether to select the triple.



  - If the triple is to be included in the summary of the document, the document-triple pair will be labeled positive, otherwise negative.
  - We need to obtain supervised labels for the triples in the training set for training the content selector. For each triple in the training set, we use ROUGE to measure the similarity to the corresponding summaries, if it is higher than a threshold then we label that triple as a positive example.
- **Rewriting**, where the selected triple are realized into natural language. The rewriter converts the selected triples into fluent summaries, where the triples serve as a content plan.



  - We train a sequence-to-sequence text generation model, similar to converting meaning-representation to natural language text.
  - The train data for this phase contains the texts and the triples extracted from them. To train the generation model, we concatenate the extracted triples from the document as the source sequence, and use the text as the target sequence.

Experiment Setting

Our main results are based on 3 news and social media summarization datasets: Gigaword, Reddit TIFU and DUC-2004. We used OLLIE, CoreNLP OpenIE and UW OpenIE as the triple extractors. We fine-tuned the RoBERTa-large as the content selector and fine-tuned the BART-large from `fairseq` as the rewriter.

Results

- **Automatic Evaluations** We report ROUGE score on the Gigaword test set and the DUC-2004 dataset, containing 1951 and 500 samples respectively.

Model	R-1	R-2	R-L
BART (2020)	37.28	18.58	34.53
BART-RXF (2021)	40.45	20.69	36.56
PEGASUS+Dot (2021)	40.60	**21.00**	37.00
OFA (2022)	39.81	20.66	37.11
ESR	**40.63**	20.62	**37.14**

Model	R-1	R-2	R-L
RT+Conv (2018)	31.15	10.85	27.68
BART (2020)	31.36	11.40	28.02
ALONE (2020)	32.57	11.63	28.24
WDROP (2021)	33.06	11.45	28.51
ESR	**33.08**	**11.52**	**28.74**

Table 1. ROUGE F1 on the Gigaword testset (left) and on the DUC-2004 dataset (right). It shows that ESR achieves or is competitive with the state-of-the-art on this dataset. **Bold** indicates the best score.

- **Modularity** To test the modularity of ESR, we report the ROUGE on Reddit TIFU reusing a rewriter trained on Gigaword in Table 2. The modularity makes it that we can train the rewriter on high resource domains and reuse it in low resource tasks. We further subsampled 1k samples from Reddit TIFU and Gigaword for training the modules to see how performance varies in the small data regime.

Model	R-1	R-2	R-L
BART (2020)	24.19	8.12	21.31
PEGASUS+Sum (2022)	29.83	9.50	23.47
BART-R3F (2021)	30.31	**10.98**	24.74
ESR			
$S_R + R_G$	**30.63**	*10.82*	**24.78**
$S_R + R_R$	29.92	10.51	24.26
$S_{R1k} + R_{G1k}$	29.67	10.09	24.00
$S_{R1k} + R_{R1k}$	29.38	10.02	23.90
$S_{R1k} + R_G$	29.09	10.07	23.86

Table 2. ROUGE F1 on R-TIFU (Reddit-TIFU). S_R means the content selector was trained on R-TIFU, R_G and R_R mean rewriter trained on Gigaword and R-TIFU respectively. 1k means that the module is trained on 1k randomly sampled subset. The content selector can be trained with low-resourced data without large dropping. **Bold** means the best and *Italics* means the best in ESR. ESR performance achieve or is compatible with the best scores. We see that training a content selector on only 1k examples and reusing the rewriter from Gigaword is on-par with using the entire Reddit TIFU.

- **Human Evaluation** We conducted a user study on Amazon MTurk where three annotators rated summaries of 100 randomly sampled texts from the Gigaword test set on faithfulness

Summaries	Sup.	Unsup.	Incoh.	Inconc.
Human-Written	96	3	0	1
BART	90	6	2	2
ESR	94	3	2	1

Table 3. Human evaluation on faithfulness. The summaries from the dataset (Human-Written) and those from ESR and the BART are annotated by 3 annotators. Crowdworkers find ESR to be more faithful than BART. The summaries are marked as supported (by the source), unsupported, or incoherent by each annotator. The final label is decided by majority vote. It is labeled inconclusive if there is no agreement.

- **Case Study** A representative case is shown in Figure 2. It shows that ESR can eliminate the hallucination and control the summarization styles with different rewriter modules.

Case Study

ST: Zairean president Mobutu Sese Seko will stay at his French Riviera residence until at least the middle of the week because of an increase in diplomatic activity, a Mobutu aide said on Sunday.

Selected Triples:
(Zairean president Mobutu Sese Seko, will stay at, his French Riviera residence)
(Zairean president Mobutu Sese Seko, will stay until, the middle of the week)

ESR: **Zairean president Mobutu** will stay at his **French Riviera residence** until the middle of week

BART: **Tanzania's** Mobutu to stay at **Riviera residence** until middle of week

Ref: Zairean president Mobutu to stay in France till mid-week

Figure 2. A case on the Gigaword testset. **ST**: source text; **Ref**: reference summary; **Selected Triples**: triples selected by the content selector. With the rewriter module trained on different dataets, the text style of ESR can be controlled. The green shows the factual correctness and the red shows the error.

Conclusion

We propose ESR, a three-phase modular abstractive summarization method. It obtains competitive performance on automatic metrics while producing more faithful summaries, and its modularity makes it have a good controllability on summary generation, and maintains a good performance on low resource data. In the future, we are adapting the ESR method to multi-document summarization datasets.

Correspondence: Shuo, Vishakh