Reducing named entity hallucination risk to ensure faithful summary generation

Eunice Akani [1,2] Benoit Favre [1] Frederic Bechet [1] Romain Gemignani [2]

[1]Aix-Marseille Univ, LIS, Marseille, France [2]Enedis, France

Problem to address

- Some text summarization systems generate unfaithful information (hallucination) regard to the source document due to pre-training on large corpora.
- We are focused on named entities, which in our use cases lead to an unfaithful summary.
- Some suggested using contrast candidate generation and selection as a post-processing method to avoid hallucination. [Chen et al., 2021]

What does this study add?

- We used a criterion called NEHR to select the summary with minimum entities hallucinated among diverse summaries generated.
- NEHR (Named Entity Hallucination Risk) is the risk of having an entity that is not faithful to the source document (entity hallucinated)
- To generate a variety of summaries we used sampling methods.

Figure 1. Unfaithful summary generated: Named entities hallucinated.

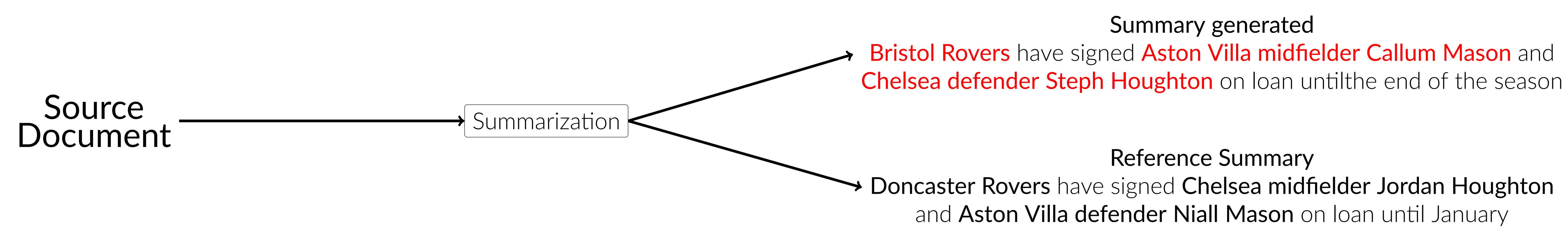
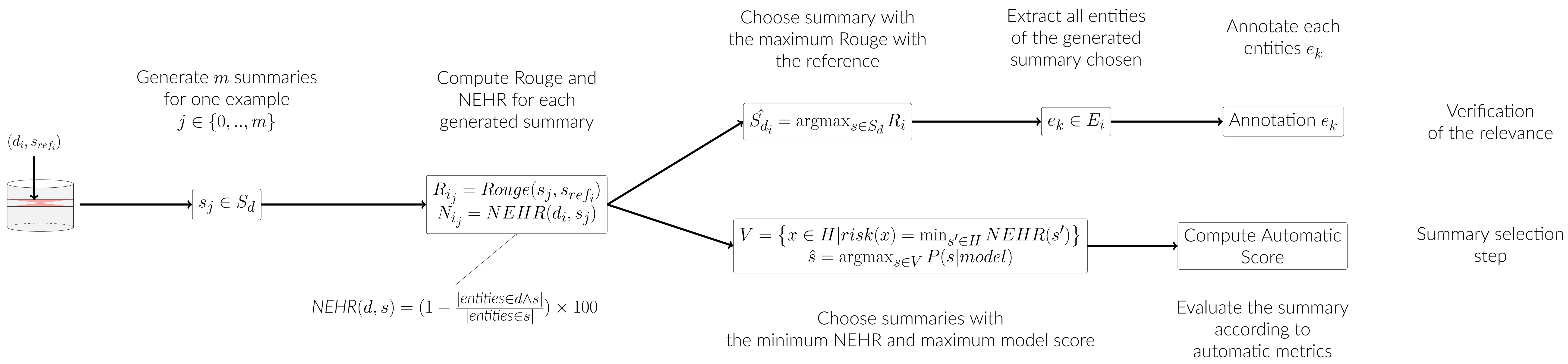


Figure 2. Proposed approach: select summaries that minimize NEHR while maximizing likelihood.



Verification of the relevance of NEHR

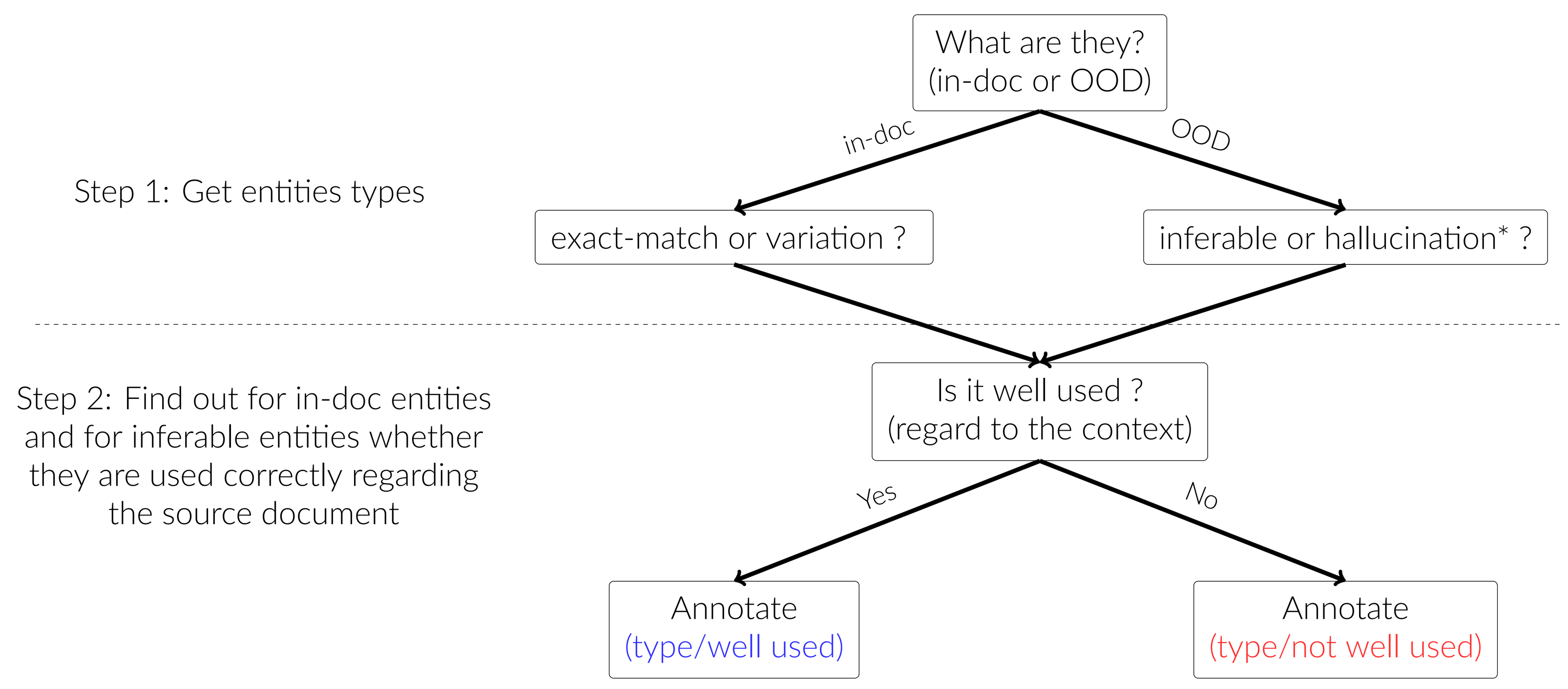
Generate multiple summaries

By varying several parameters such as:

- Temperature $\in [0.5, 0.6, 0.7, 0.8, 0.9]$;
- Top k $\in [40, 50, 60]$;
- Top p $\in [0.75, 0.80, 0.85, 0.90, 0.95]$.

$\Rightarrow$ 75 summaries + beam + greedy = 77 summaries for each examples

Entity Annotation



Annotation Results

- 3 annotators on 50 generated summaries randomly selected from the test set of CNN/DM following the annotation process.

	in-document		out-document	
Entity dist. (%)	79.7		20.3	
Type	*exact.*	*var.*	*inf.*	*hall.*
Type dist. (%)	62.8	37.2	28.8	71.2
% correct	90	90	88	-

Table 1. % of correctly used entities for each subset of in-document and out of document entities.

Summary selection

Selection criterion

It based on both NEHR and model scores to select the summary with a minimum hallucinated entities :

$$V = \left\{ x \in H | risk(x) = \min_{s' \in H} NEHR(s') \right\} \quad (1)$$

$$\hat{s} = \underset{s \in V}{\text{argmax}}\, P(s|model) \quad (2)$$

Dataset and Model

- **CNN/DM**: An abstractive text summarization dataset based on the CNN articles and the DailyMail websites.
- **XSum**: A more abstractive text summarization dataset based BBC articles.
- **BART**: A language model used to perform text generation including automatic text summarization.

Results on XSUM et CNN/DM

	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	NEHR ↓	%HallSum ↓
Beam 4	**43.74**	**20.84**	**30.44**	0.5	3.86
Best Proba	41.99	18.96	28.01	2.6	20.57
Entailment	43.61	19.69	29.26	1.62	12.92
our	42.19	19.12	28.24	**0.003**	**0.035**

Table 2. Results on CNN/DM

	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	NEHR ↓	%HallSum ↓
Beam 4	**45.32**	**22.20**	**37.10**	27.67	52.48
Best Proba	40.26	16.79	31.29	31.05	61.24
Entailment	40.92	17.14	31.96	27.08	54.98
our	40.16	16.54	31.31	**6.92**	**21.49**

Table 3. Results on XSUM

Conclusion

- Our study shows that NEHR can be used as selection criterion combine to model score. It gives competitive ROUGE score on CNN/DM and drops dramatically the hallucination risk on XSum.
- Human evaluation on XSum summaries shows that the occurring entities were more often correct with respect to those obtained without our selection criteria.

Aix-Marseille Univerty and Enedis **eunice.akani@lis-lab.fr**