Unsupervised Token-level Hallucination Detection from Summary Generation By-products

Andreas Marfurt
Idiap Research Institute, Switzerland
EPFL, Switzerland
andreas.marfurt@idiap.ch

James Henderson
Idiap Research Institute, Switzerland
james.henderson@idiap.ch

Abstract

Hallucinations in abstractive summarization are model generations that are unfaithful to the source document. Current methods for detecting hallucinations operate mostly on noun phrases and named entities, and restrict themselves to the XSum dataset, which is known to have hallucinations in 3 out of 4 training examples (Maynez et al., 2020). We instead consider the CNN/DailyMail dataset where the summarization model has not seen abnormally many hallucinations during training. We automatically detect candidate hallucinations at the token level, irrespective of its part of speech. Our detection comes essentially *for free*, as we only use information the model already produces during generation of the summary. This enables practitioners to jointly generate a summary and identify possible hallucinations, with minimal overhead. We repurpose an existing factuality dataset and create our own token-level annotations. The evaluation on these two datasets shows that our model achieves better precision-recall tradeoffs than its competitors, which additionally require a model forward pass.

1 Introduction

Large pretrained Transformers (Vaswani et al., 2017; Devlin et al., 2019) have considerably advanced the state of the art in abstractive summarization (Liu and Lapata, 2019; Lewis et al., 2020; Zhang et al., 2020). However, model hallucinations – where the information in the generated summary is not faithful to the source document – are a prominent remaining failure mode of these models.

A lot of recent work has addressed this problem, predominantly on the XSum dataset (Narayan et al., 2018). XSum is an outlier, however, in that over 75% of its reference summaries contain hallucinations (Maynez et al., 2020). Models trained (or finetuned) on this dataset are consequently prone to hallucinate themselves when summarizing an article. Additionaly, current work focuses on detecting hallucinations for noun phrases and named entities (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021), sometimes with the addition of dates and numbers (Narayan et al., 2021). Recent work has shown, however, that summarization models also make mistakes in other parts of speech, such as predicates (Pagnoni et al., 2021).

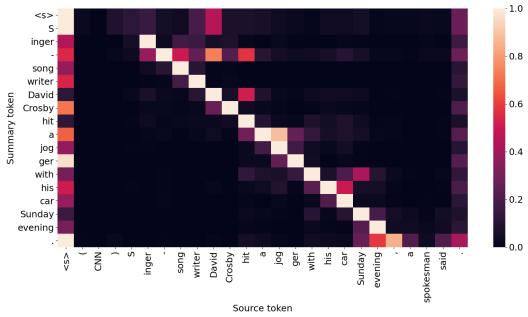


Figure 1: BART cross-attentions align copied segments of the summary with the respective segments in the source. Attention weights are normalized by row. Only the first summary and source sentences are shown.

In this paper, we aim to expand the current line of research to a different dataset, and to remove the restriction to entities. We use the diagonal cross-attention patterns present in Transformer-based abstractive summarization models (see Figure 1) to align the summary with the source document. We detect hallucinations in an unsupervised fashion for segments of aligned and unaligned tokens by computing statistics from the encoder's self-attentions and the decoder's next-word probabilities. These by-products arise when generating a summary with any Transformer model. In this paper, we use BART (Lewis et al., 2020). We evaluate our approach on two datasets.[1] We repurpose the factuality dataset FRANK (Pagnoni et al., 2021), but only 0.4% of tokens turn out to be hallucinations. Therefore, we additionally create our own dataset

[1]Our data and code are available at https://github.com/idiap/hallucination-detection.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 248 - 261

called TLHD-CNNDM, which contains token-level annotations on examples heuristically selected to have a higher chance of containing a hallucination. Indeed 14.2% of tokens in TLHD-CNNDM are hallucinations. Our method demonstrates good results compared to its competitors, while at the same time requiring negligible additional computation. At the same time, hallucination detection proves to be a difficult task, in particular on intrinsic hallucinations (defined in Section 3), where all models struggle to detect any hallucinations.

2 Related Work

Several different methods have been proposed to detect hallucinations. Specialized decoding strategies are used to nudge the model to stay closer to the source vocabulary (Aralikatte et al., 2021) or its entities (Narayan et al., 2021). Multiple studies use automatic question generation and answering models to ask questions about entities in the generated summary, and try to answer them from the source document (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021). If the question cannot be answered from the source document, the entity is considered a hallucination. Filippova (2020) determine the degree of hallucination from the differences in probabilities assigned by a conditional and an unconditional language model. In the related area of factuality detection, Cao et al. (2022) use the same idea to identify hallucinated but factual summaries. Entailment-based classifiers are used to evaluate a summary's factuality at the level of text or dependency arcs (Falke et al., 2019; Goyal and Durrett, 2020). It is also common to create synthetic data for a classifier by corrupting the input, for hallucinations (Zhou et al., 2021) as well as factuality (Cao et al., 2020; Kryściński et al., 2020). However, the error distributions obtained synthetically can differ from those of models (Goyal and Durrett, 2021). More types of factuality errors are identified in Pagnoni et al. (2021) with a detailed human annotation, finding discourse and semantic frame errors. These detection methods can be used to identify mistakes or rerank multiple outputs (e.g. Ladhak et al., 2022).

3 Hallucination Detection

Definition. We adopt the definition from Maynez et al. (2020), and define *intrinsic hallucinations* as combinations of information from the source document that cannot be inferred from it, and *extrinsic hallucinations* as information that is not present in the source document. Paraphrases and information that can be directly inferred from the source document, however, do not constitute hallucinations. Furthermore, whether some information is a hallucination is an orthogonal problem to whether that information is factually correct, a question we do not consider in this paper.

3.1 Unsupervised Hallucination Detection

In the process of generating a summary, a Transformer-based abstractive summarization model creates a number of by-products, such as decoder next-token generation probabilities, encoder and decoder self-attentions, and decoder to encoder cross-attentions, for each layer and attention head of the model. These can be easily accessed from e.g. the HuggingFace transformers library (Wolf et al., 2020).

Motivation. It is debated whether model attentions can be used to explain model decisions (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) and how much a Transformer encoder's output representation still represents the token at its position in the input (Brunner et al., 2020). Nevertheless, we posit that the diagonal attention patterns observed in Figure 1, together with the fact that the source and target tokens match for the entire segment, is a strong enough signal to claim that a summarization model copied this segment from the source.

Additionally, we conjecture that the faithfulness of a summary to the source document is not inherently a question that spans multiple sentences, in contrast to a summary's factuality (Pagnoni et al., 2021). As a consequence, we detect hallucinations at the token level by processing summary sentences in isolation.

Initial alignment. From the observations above, we start by aligning summary and source positions based on cross-attentions. In BART cross-attentions, the maximum cross-attention weight is often put on the beginning-of-sequence token in the source. If the token is a preposition, a high attention weight is also put on its preceding and succeeding tokens. We therefore accept a target-source alignment of target token t_i iff it matches a source token in its top-4 cross-attention weights. This constitutes our initial alignment.

Context voting. In a second step, we expand the initial alignment with a position-based voting algo-

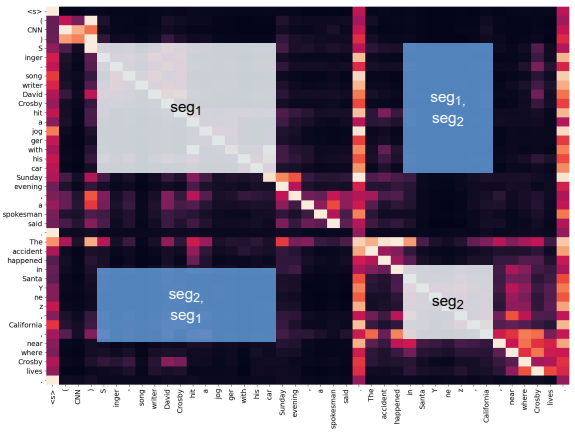


Figure 2: BART encoder self-attentions relate the aligned segments seg_1 and seg_2 of the source document (grey boxes) by their interactions (blue boxes). Only the first two source sentences are shown.

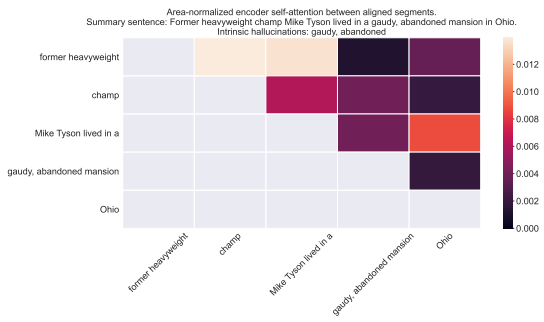


Figure 3: Association strength α between aligned segments. The intrinsic hallucinations in the fourth segment show the least interaction with other segments. Full example in Appendix B.

rithm. For each target token t_i, its context tokens $t_{i-l}, \ldots, t_{i-1}, t_{i+1}, \ldots, t_{i+l}$ in a window of size l[2] around t_i vote on the expected source position of t_i given their own alignment and an assumed diagonal attention pattern. If a token is not aligned to the source, it does not vote. We accept a vote when at least half the neighboring tokens agree. We perform voting for a maximum of 10 rounds, and we stop early when it has converged, which often happens after 2 rounds.

After these two alignment stages, we have a set of aligned segments, with a token-level correspondence between summary and source, and a set of unaligned tokens. We now look to detect intrinsic hallucinations in the former set, and extrinsic ones in the latter.

Classifying aligned tokens. Aligned tokens appear in the source document, and consequently do not constitute extrinsic hallucinations. To assign a probability of them being intrinsic hallucinations, we compare characteristics of their aligned source segments. Maynez et al. (2020) speculate that intrinsic hallucinations are potentially a failure of document modeling. We add that the encoder may also have performed well at document modeling, but the communication to the decoder through the representational bottleneck may have failed. In the latter case, we should be able to read the association of two source segments from the strength of the encoder's self-attentions between the two segments. We determine the association strength α of two aligned segments seg_1 and seg_2 by the area-normalized sum of encoder self-attention weights (enc_{ij} and enc_{ji}) between the two segments:

$$\alpha(\text{seg}_1, \text{seg}_2) = \frac{\sum_{i \in \text{seg}_1, j \in \text{seg}_2} \text{enc}_{ij} + \text{enc}_{ji}}{2 * |\text{seg}_1| * |\text{seg}_2|} \quad (1)$$

where i and j are the source indices of segments seg_1 and seg_2, and $|.|$ is the cardinality. Figure 2 visualizes the areas whose attention weights are summed with blue boxes. The score for a segment is the mean α to all other segments in its summary sentence. The higher the score, the higher our confidence in the two segments being semantically close, and therefore not intrinsic hallucinations. As an example, Figure 3 shows that the fourth segment has the smallest association strength to the other segments. Indeed, this is an intrinsic hallucination. It talks about the present state of the mansion, while the predicate concerns the past.

Classifying unaligned tokens. While unaligned tokens can still appear in the source document and result in an intrinsic hallucination, the prevalent error mode for this set of tokens are extrinsic hallucinations. We found that generated summaries sometimes contain sentences entirely unrelated to the article, most likely an artefact of data collection. Our first score β_{align} is the fraction of the summary sentence tokens that are aligned.

For unaligned tokens in mostly-aligned sentences, we conjecture that generations by a strong language model fit in well (both syntactically and semantically) with the source document and the summary written so far, and thus should be expected by the model. In the opposite case, unexpected generations lead to a higher amount of surprisal. The expected surprisal of a language model can be quantified with the entropy of its next-word

[2]We choose $l = 3$ as our window size.

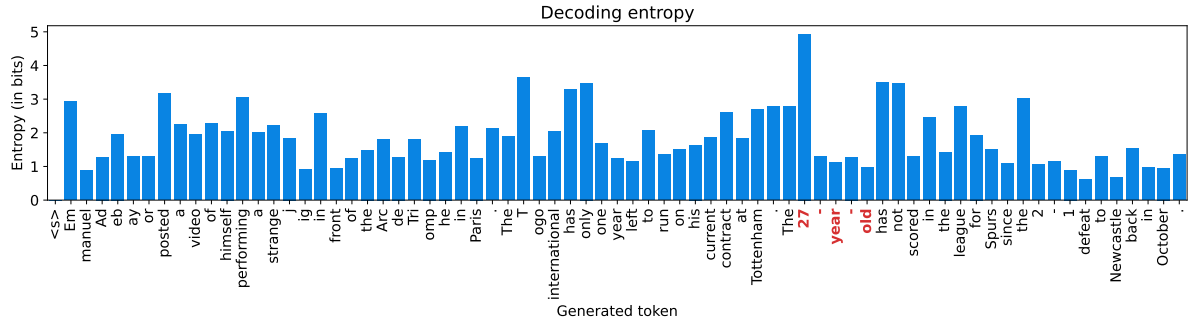


Figure 4: Summary containing an extrinsic hallucination (tokens in bold red). The decoding entropy of the first hallucinated token is high, those of the subsequent tokens are low. We determine the hallucination score (Eq. 2) of the entire segment (Definition in Eq. 3) from its first token.

decoding probabilities (Meister et al., 2020). Figure 4 shows the decoding entropy of an example summary. We thus propose a second score β_{entropy} as the inverse smoothed decoding entropy:

$$\beta_{\text{entropy}}(t_i) = \frac{1}{H(t_i) + 1} \tag{2}$$

with $H(t_i)$ the entropy of the next-word probability distribution of target token t_i.

Only the generation of the first token of an unexpected segment is surprising (as seen in Figure 4), and subsequent completions of the segment have high probability and low entropy. We therefore split a span of unaligned tokens into segments based on the decoding entropy. The construction is as follows: As long as the decoding entropy of the next token t_i decreases the mean decoding entropy of the current segment seg, it is added. Otherwise a new segment is started.

$$\text{seg}' := \begin{cases} \text{seg} \cup t_i & \text{if } H(t_i) < \frac{\sum_{t_j \in \text{seg}} H(t_j)}{|\text{seg}|}, \\ t_i & \text{otherwise.} \end{cases} \tag{3}$$

Converting scores to probabilities. All our faithfulness scores are nonnegative, and upper bounded by 1. A higher score means less chance of hallucination. We therefore convert each faithfulness score s to a hallucination probability p by scaling and inverting it.

$$p = 1 - \frac{s - s_{\text{min}}}{s_{\text{max}} - s_{\text{min}}} \tag{4}$$

where s_{min} and s_{max} are the minimum and maximum scores across the entire dataset. In an offline evaluation setting, one can compute all scores on a dataset first, and then get s_{min} and s_{max}. For the online setting, these values have to be set. On our two datasets, we observe that the minimum and maximum values do not change much, so we expect the current values to transfer to new datasets. They are $[0, 0.02]$ for α, $[0.08, 0.71]$ for β_{entropy}, and β_{align} is already in the correct range.

BART-GBP. As we will see in the ablation study in the results in Section 5, the association strength α decreases the performance of our detection method. Our final model, *BART-GBP* (BART generation by-products), therefore only uses the β_{align} and β_{entropy} scores.

4 Experiments

We study CNN/DailyMail (Hermann et al., 2015), a summarization dataset known to be highly extractive (Grusky et al., 2018) and therefore less likely to contain a lot of hallucinations.

4.1 Datasets

Finding an existing dataset to evaluate our method is difficult, since we need access to the model's attentions and decoding probabilities alongside the outputs.

FRANK. We repurpose FRANK, a factuality metric evaluation dataset (Pagnoni et al., 2021). It consists of 250 summaries from the CNN/DailyMail test set, obtained from SummEval (Fabbri et al., 2021). FRANK introduces a typology of factual errors, which we convert to hallucination annotations by using examples of predicate, entity and circumstance errors as candidates for intrinsic hallucinations, and out-of-article errors as candidates for extrinsic hallucinations. Our publically available model version produces slightly different outputs from theirs, so we manually correct labels

where the outputs differ. Our adapted dataset contains 57 hallucinated words (31 intrinsic, 26 extrinsic) which corresponds to 0.4% of the 15,700 total words. At the sentence level, 3.5% contain at least one hallucinated word (31/897), while at the summary level it is 9.2% (23/250).

TLHD-CNNDM. Since the number of hallucinations in FRANK is low, we additionally collect human annotations ourselves. We produce BART model outputs for the CNN/DailyMail test set (excluding the FRANK examples) by using the standard HuggingFace implementation with the default parameters. To arrive at an interesting dataset, we first rank summary sentences by two criteria: 1) the number of non-contiguous alignments to the source document found by lexical overlap, and 2) the number of words that do not appear in the source document. Both criteria are length-normalized. We pick the top 75 examples from both lists, arriving at 150 summary sentences. We then perform a human annotation as detailed in Appendix C. Our dataset contains 299 hallucinated words out of a total 2,100 (14.2%). Of those hallucinations, 51 are intrinsic, and 248 are extrinsic. Of the 150 sentences, 78 contain at least one hallucination (52%). The annotator agreement with the majority class (following Durmus et al., 2020) is 94.6%, and 73.9% and 86.3% for intrinsic and extrinsic hallucinations, respectively. We name our dataset *TLHD-CNNDM* (token-level hallucination detection for CNN/DailyMail).

4.2 Model Details

For generating our summaries, attentions and decoding probabilities, we use the BART-large model finetuned on CNN/DailyMail (`'facebook/bart-large-cnn'`) from the HuggingFace transformers library[3], with its default parameters. In generation with beam search, multiple beams are active at each generation step, but only one beam is eventually selected. We extract the attention and decoding probabilities of this beam with our own code. When inspecting cross-attentions, we found layers 10 and 11 (out of 12) to show the cleanest diagonal patterns (as presented in Figure 1). Other layers either have less focused attention, or they look at the previous token (mostly lower layers), the beginning-of-sequence token, or periods. We average the attentions from layers 10 and 11. We select the same layers for the encoder self-attentions.

[3] https://github.com/huggingface/transformers

4.3 Baselines

As baselines, we use four classes of models: lexical overlap, an entity-focused question-generation-answering model, a dependency entailment-based model, and a token-level classification model trained on synthetic data.

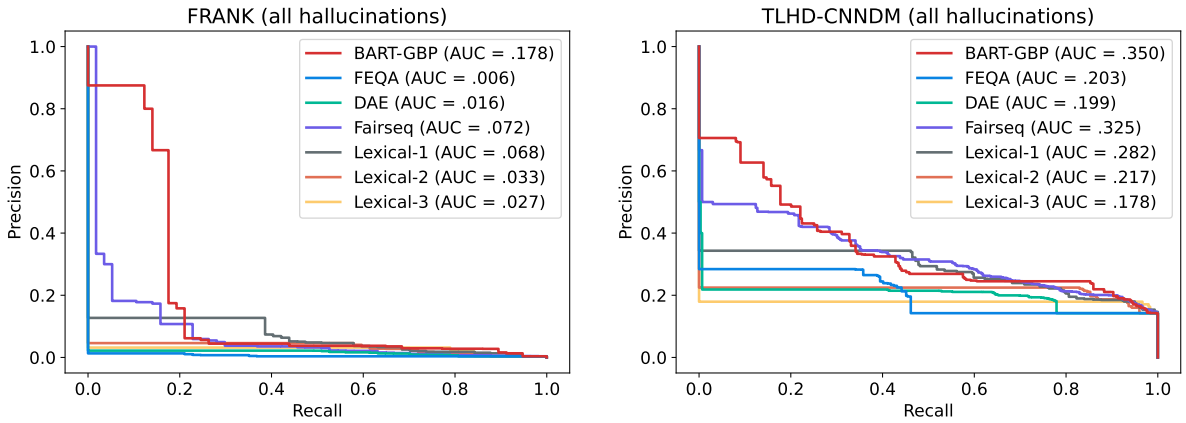
Lexical-n. This baseline lexically aligns the summary and the source document. It greedily adds the longest matching span, down to a span length of n. This baseline classifies all unaligned tokens as (presumably extrinsic) hallucinations. For aligned tokens, our most successful heuristic determines the hallucination probability for each aligned span as the fraction of aligned tokens that have an alignment in the same source sentence as the current span:

$$1 - \frac{|\text{tokens aligned to same source sentence}|}{|\text{all aligned tokens}|}. \quad (5)$$

FEQA. FEQA (Durmus et al., 2020) generates questions about the summary's entities, then tries to answer them from the source document. It then computes the token-level F1 score between the summary's text and the predicted text span from the source. Unmatched answers indicate hallucinations. We compute word-level probabilities by averaging the F1 scores of all spans the word is part of.

DAE. Dependency arc entailment (DAE) (Goyal and Durrett, 2020, 2021) decides from its dependency arcs whether the generated summary sentence is entailed by the source document. While DAE is technically a factuality detection method, we conjecture that hallucinations in the summary should not be entailed by the source document either. In footnote 6 of Goyal and Durrett (2021), the authors propose that a word is non-factual if any of its arcs is non-factual. We therefore compute word hallucination probabilities as the maximum probability of non-factuality of its dependency arcs. We use their model variant trained with entity-based synthetic data on CNN/DailyMail.

Fairseq. With the help of synthetic training data, where factual tokens have been automatically replaced with hallucinations, pretrained language models can be finetuned to directly predict a hallucination label for each input token



(a) FRANK

(b) TLHD-CNNDM

Figure 5: Precision-recall curves for all hallucinations in the FRANK and TLHD-CNNDM datasets.

Method	Best F1	PR AUC	ROC AUC
FRANK			
FEQA*	0.0245	0.0062	0.3327
DAE*	0.0419	0.0157	0.7164
Fairseq w/o ref*	0.1651	0.0723	0.8129
Fairseq w/ ref*	0.0682	0.0232	0.7017
Lexical-1	0.1913	0.0677	0.8788
Lexical-2	0.0854	0.0335	0.8058
Lexical-3	0.0610	0.0268	0.7672
BART-GBP	**0.2778**	**0.1777**	**0.8934**
TLHD-CNNDM			
FEQA*	0.3156	0.2031	0.3899
DAE*	0.3167	0.1988	0.5803
Fairseq w/o ref*	**0.3957**	0.3255	**0.7375**
Fairseq w/ ref*	0.2672	0.1714	0.5521
Lexical-1	0.3937	0.2819	0.6846
Lexical-2	0.3535	0.2166	0.4802
Lexical-3	0.3025	0.1785	0.2599
BART-GBP	0.3806	**0.3502**	0.7332

Table 1: Best F1 score on the precision-recall curve, area under precision-recall curve, and area under the ROC curve. Methods marked with * require an additional model forward pass, which increases runtime and resource use.

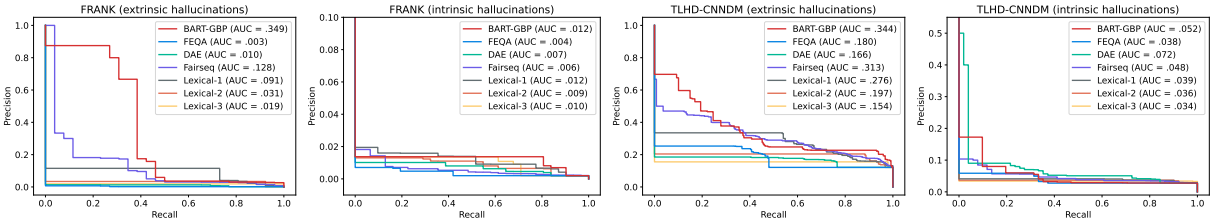
(Zhou et al., 2021). We use the model finetuned on XSum and evaluate how it transfers to the CNN/DailyMail dataset. Since we compare to our unsupervised method, we leave retraining the model on CNN/DailyMail to future work. We evaluate both model settings, with and without access to the reference summary. We call this method *Fairseq* based on its Github repository name.

5 Results

We use precision-recall curves to evaluate the hallucination detection methods. Precision-recall is the preferred metric when finding the instances of the positive class (hallucinations) has exceptionally high value compared to the instances of the negative class. Appendix A also shows ROC curves.

Main result. Our main result is shown in Table 1, which considers performance when classifying hallucinations of both intrinsic and extrinsic type. We present the best F1 score on the precision-recall curve, the area under the precision-recall curve, and the area under the ROC curve. Additionally, we show whether the method requires an additional model forward pass, which incurs a longer runtime and higher resource costs, by marking the respective methods (with *). BART-GBP performs best on the FRANK dataset, and has the largest AUC for precision-recall on the TLHD-CNNDM dataset. For the other metrics, it is close behind the highest score, all while being completely unsupervised. Fairseq without access to the reference summary performs well on TLHD-CNNDM, but worse on FRANK. The setting without access to the reference summary does better across all datasets and metrics, and is therefore reported from now on.

The precision-recall plots in Figure 5 give further details on the main result. BART-GBP manages to get high precision for the data points where it is most certain, something other methods struggle with. At higher levels of recall, the difficulty of the task leads to lower precision scores across all methods. The FRANK dataset, where only 0.4% of tokens are hallucinations, is very challenging (see Figure 5a). With 14.2% of positive labels, TLHD-CNNDM is less extreme, but still proves to be difficult for all methods, as seen in Figure 5b.



(a) Extrinsic hallucinations (b) Intrinsic hallucinations (c) Extrinsic hallucinations (d) Intrinsic hallucinations

Figure 6: Precision-recall curves for the label subsets of extrinsic and intrinsic hallucinations in the FRANK (6a, 6b) and TLHD-CNNDM (6c, 6d) datasets.

Extrinsic hallucinations. Figures 6a and 6c show the models' performance on the label subset of extrinsinc hallucinations. To evaluate on this subset, we remove data points that are gold intrinsic hallucinations in order to not unfairly penalize models for detecting those, and vice versa for evaluation of intrinsic hallucinations. Apart from BART-GBP and Fairseq, the Lexical-1 baseline manages to find some hallucinations. However, it does not provide a fine-grained trade-off between precision and recall, in contrast to BART-GBP.

Intrinsic hallucinations. As we can see from Figures 6b and 6d, finding intrinsic hallucinations proves to be very difficult for all methods. We therefore zoom in both graphs on the y-axis. BART-GBP performs well relative to the baselines. Notably for the TLHD-CNNDM dataset, DAE manages to find some hallucinations at some of its highest probability selections, but quickly diminishes at higher recall.

In summary, BART-GBP gets consistent and very competitive results in both datasets and on all label subsets, even while being an unsupervised method. The ROC curves in Figures 8 and 9 in Appendix A further confirm this finding.

Ablation study. We are interested to see how each of our designed scores contributes to finding hallucinations. In Table 2, we show an ablation study with the area under the precision-recall curve as the performance metric. We see that of all individual scores, β_{align} performs best. Combining it with β_{entropy} (by taking the maximum of both probabilities for each token) further improves results on the TLHD-CNNDM dataset, but not on FRANK. α performs barely above a baseline that would classify all data points as hallucinations. This came as a surprise to us, as we expected α to perform better from the motivation in Section 3.1. Adding α to the β scores decreases performance drastically.

Scores	FRANK	TLHD-CNNDM
α	0.0051	0.1440
β_{align}	**0.1993**	0.3260
β_{entropy}	0.0685	0.3198
β_{align}, β_{entropy}	0.1777	**0.3502**
α, β_{align}, β_{entropy}	0.0390	0.1687

Table 2: Ablation study for different combinations of scores. Metric is area under precision-recall curve. BART-GBP is the combination of β_{align} and β_{entropy}.

This comes from the fact that our scores are not calibrated, so the distribution of each score will be different. As a result, when taking the max of multiple scores, one of them may dominate. When we plot a histogram of our scores' values, we see that this is the case for α, leading to such a performance deterioration in the case of combining all three scores. Since α on its own does not score well, we do not further calibrate our scores.

Maximum possible hallucination recall. We motivated our approach by arguing that token-level methods are superior to entity-based question-generation-answering systems (like FEQA) or dependency arc entailment-based DAE. These methods may miss some hallucinated tokens as they only compute hallucination probabilities for a subset of all tokens. To verify how many these are, we analyze the recall each method achieves when it classifies all tokens that it considers as positives.

The results are shown in Table 3. The disadvantage for FEQA and DAE is substantial. FEQA classifies less than half of the tokens labeled as hallucinations in the FRANK and TLHD-CNNDM dataset. DAE is limited to a recall of around 80%, as it cannot detect tokens that are not part of one of the dependency arcs considered for entailment.

Method	Maximum possible recall
FRANK	
FEQA	38.60%
DAE	80.70%
Fairseq	100.00%
Lexical-n	100.00%
BART-GBP	100.00%
TLHD-CNNDM	
FEQA	46.15%
DAE	77.93%
Fairseq	100.00%
Lexical-n	100.00%
BART-GBP	100.00%

Table 3: Maximum possible recall of FEQA (entity-based), DAE (dependency arc entailment), and the token-level methods Fairseq, Lexical-n and BART-GBP.

Score	All	Extrinsic	Intrinsic
FRANK			
Aligned (α)	50.88%	11.54%	83.87%
Unaligned (β_{entropy})	52.63%	96.15%	16.13%
Both (β_{align})	100.00%	100.00%	100.00%
TLHD-CNNDM			
Aligned (α)	19.06%	11.29%	56.86%
Unaligned (β_{entropy})	81.27%	88.71%	45.10%
Both (β_{align})	100.00%	100.00%	100.00%

Table 4: Maximum possible recall of aligned and unaligned token scores wrt. all, extrinsic, or intrinsic hallucinations.

Maximum recall of (un)aligned tokens. Aligning the summary with the source document forms the basis of our method. How many hallucinations are part of aligned spans, and how many are unaligned? We perform this analysis in Table 4. We can see that extrinsic hallucinations are mostly part of unaligned spans, which are scored by β_{entropy}. Intrinsic hallucinations in the FRANK dataset are mostly part of aligned spans, scored by α. In the TLHD-CNNDM dataset, however, intrinsic hallucinations are only part of aligned spans around half of the time.

Note that aligned and unaligned scores can add up to slightly more than 100%. This occurs when some BPE tokens of the same word are aligned and others are not (e.g. when a name appears together with a possessive *'s*).

6 Conclusion

We have presented BART-GBP, a method to detect hallucinations from the by-products of summary generation of a BART abstractive summarization model, trained and evaluated on CNN/DailyMail. We first aligned the segments of the summary and source document using cross-attentions, and then used encoder self-attentions and decoding probabilities to detect intrinsic and extrinsic hallucinations, respectively. This happens with minimal computational overhead, compared to prior work that uses external models that require an additional model forward pass. Our evaluations show that this is a difficult task, and especially intrinsic hallucination detection needs to be addressed by future work. We hope to contribute to this endeavor with our method and our token-level annotated dataset, TLHD-CNNDM.

Limitations

The results in this paper are limited by several factors. Firstly, the definition of what constitutes a hallucination is neither set in stone, nor a mathematical construct, and therefore open to interpretation. We experienced this first-hand from the feedback of our annotators. This makes the task of teaching a model to identify hallucinations all the more difficult, and the gap to optimal performance in the results (for all methods) makes this visible.

Another limitation is given by the model under study. We already mentioned in Section 3.1 that the interpretability of attention patterns is a debated topic in the research community. A model trained to faithfully *explain* its decisions would be even better suited to perform this kind of analysis.

Transfer to other models. While we do not assume that our method transfers easily to some attention-based RNN architectures, we saw indications that it could transfer to other Transformer-based summarization models. In initial experiments, we have used BERTSUMABS (Liu and Lapata, 2019), which shows very similar cross-attention patterns (see Figure 7). There are some small differences, however. BERTSUMABS puts its maximum attention weight to the copied word more often, but still shows a lot of attention to CLS/SEP tokens in the source and BOS/EOS tokens in the summary. Additionally, the tokenization is different which can have an impact on the alignment stage. In BART, for example, the same word can

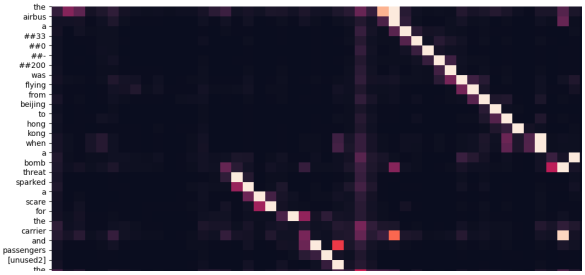


Figure 7: BERTSUMABS cross-attention patterns are very similar to those of BART, both Transformer-based summarization models.

be tokenized in different ways when it is preceeded by the BOS token, a whitespace, or punctuation. This sometimes prevented our method from aligning the same word due to unmatched tokens.

Transfer to other datasets. We do not expect these results to transfer to datasets that have a large percentage of hallucinations, i.e. XSum. We are not aware of other datasets with those same hallucination characteristics. However, we expect that other summarization datasets could benefit from our method, especially those that are similarly extractive as CNN/DailyMail. The scoring range to convert scores into probabilities may have to be recomputed.

Prevalence of sports topics in hallucinations. The prevalence of sports topics in CNN/DailyMail hallucinations hints at divergence issues between the source and reference (Wiseman et al., 2017; Dhingra et al., 2019; Kryscinski et al., 2019) for these topics: True additional information (such as standings) is added by the author/editor. It is interesting to note that while models trained on XSum learn to hallucinate consistently, CNN/DM models learn to hallucinate on sports topics. While removing hallucinations from the training data could address hallucinations, this seems infeasible, and detecting hallucinated model outputs is a more practical approach.

Ethical Considerations

By using a large pretrained language model, this study inherits the issues that come with these models, i.e. reproduction of biased or offensive content that appeared in the pretraining corpus, which includes documents on the web. Unexpected and unwanted model behavior should be reduced. Detecting hallucinations is one of the methods to do so, which can prevent misrepresentation of the text to be summarized by the model, and avoid distributing potentially misleading and in the worst case harmful content. On the other hand, a danger in using an imperfect model to detect hallucinations can be to create a false sense of security and lower the vigilance of people tasked with checking model outputs.

In this study, we also conducted a human evaluation. The privacy of our annotators is respected by labeling each example's answers with annotator_0, annotator_1 and annotator_2, respectively. Their answers consist exclusively of an extracted text span from the summary sentence in question. No personal information was collected. With regard to the presented content in the evaluation, the articles are part of the publically available CNN/DailyMail test set, and supposedly do not contain offensive content. The generated summaries were checked manually. We did not hear any negative feedback from our annotators in this or any other regard.

Acknowledgments

This work was supported as a part of the grant Automated interpretation of political and economic policy documents: Machine learning using semantic and syntactic information, funded by the Swiss National Science Foundation (grant number CRSII5_180320). We would also like to thank our family, friends and colleagues who graciously volunteered to annotate the TLHD-CNNDM dataset.

References

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6078–6095.

Gino Brunner, Yang Liu, Damian Pascual Ortiz, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, 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*, pages 1693–1701.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Wojciech Kryscinski, 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. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen Mckeown. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

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.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammed Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A ROC Results

Figures 8 and 9 show the ROC curves on the FRANK and TLHD-CNNDM datasets. There is a large label imbalance in both datasets, with the positive class only making up 0.4% of FRANK's labels, and 14.2% of those in the TLHD-CNNDM dataset. This has to be considered when looking at these figures.

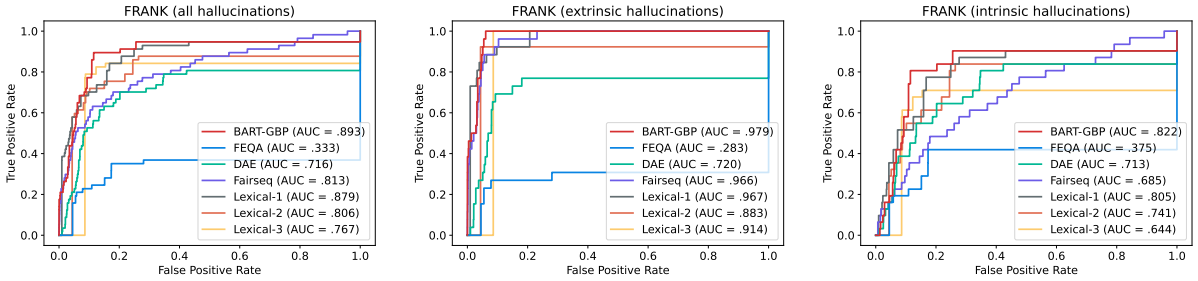
BART-GBP performs best on both datasets and label subsets, except for intrinsic hallucinations in the TLHD-CNNDM dataset in Figure 9c, where DAE and Lexical-1 perform better.

One thing that is easily visible from the ROC curves is the fraction of positive labels that can be discovered by a detection method. When a curve flattens out, it is no longer able to find more hallucinations without labeling all tokens as positive. This further highlights the strengths of the token-level methods BART-GBP and Lexical-n.

B Hallucination Examples

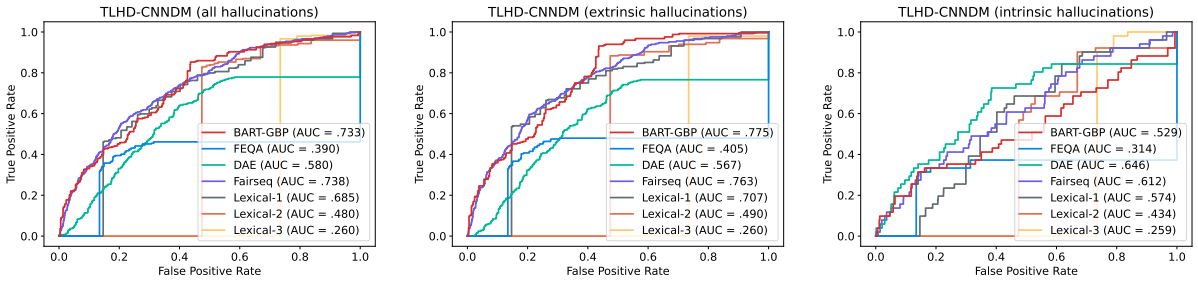
We present two examples of hallucinations, one of intrinsic hallucination from the FRANK dataset, and one of extrinsic hallucination from the TLHD-CNNDM dataset. In the former example, Mike Tyson's mansion is now in a gaudy, abandoned state, but was not while he still lived in it. In the latter example, the name of the stadium (Old Trafford) is never mentioned in the article, so it is an extrinsic hallucination. As an aside, factuality cannot be determined, since the article only talks about a "meeting" of the two teams and does not mention the home team.

Intrinsic hallucination from FRANK.
Article: (CNN)A trip to a former heavyweight champ's gaudy, abandoned mansion. The tallest



(a) All hallucinations (b) Extrinsic hallucinations (c) Intrinsic hallucinations

Figure 8: ROC curves for hallucinations in the FRANK dataset.



(a) All hallucinations (b) Extrinsic hallucinations (c) Intrinsic hallucinations

Figure 9: ROC curves for hallucinations in the TLHD-CNNDM dataset.

and fastest "giga-coaster" in the world. A dramatic interview with a famed spiritual leader – and the tearful reaction by one of his former students. These are some of the best videos of the week: In the 1980s and '90s – before he moved to Vegas and started keeping tigers as pets – former heavyweight boxer Mike Tyson lived in a Southington, Ohio, mansion. The home featured an indoor swimming pool, a marble-and-gold Jacuzzi (with mirrored ceiling, naturally) and an entertainment room large enough for small concerts. Tyson sold the house in 1999; it's due to become, of all things, a church. The video can be seen at the top of this story. Not a fan of roller coasters? You may want to skip the next video – but for the rest of us, the thrill of watching is the next best thing to being there. The Fury 325 can be found at Carowinds amusement part in Charlotte, North Carolina. Watch the video: In a CNN exclusive, Alisyn Camerota looked into allegations that Bikram yoga creator Bikram Choudhury sexually assaulted six former students. "He's a person who's based a lot of truths on a lot of lies," said Sarah Baughn, who alleges that Choudhury sexually assaulted her. Watch the video: CNN's Karl Penhaul spoke to a shepherd who witnessed the final seconds of Germanwings Flight 9525, which crashed in the French Alps last week. "I saw the plane heading down along the valley and I said, 'My God, it's going to hit the mountain,' " Jean Varrieras told Penhaul. "I ducked my head. ... Then after that, I saw the smoke." Watch the video: Magician and comedian Penn Jillette was part of a panel speaking to CNN's Don Lemon about the controversial Indiana religious freedom law. Jillette, an avowed atheist and libertarian, noted "we are not talking about forcing people to engage in gay sex, or even endorse gay sex." His provocative opening led to an energetic back-and-forth with the Alliance Defending Freedom's Kristen Waggoner and the ACLU's Rita Sklar. Watch the video: A professor of physics at a British university asked 100 people to create a composite with facial features they thought were beautiful – and then asked another 100 to rate their attractiveness. You'll never guess what celebrities best fit the model. Watch the video:

BART summary: Former heavyweight champ Mike Tyson lived in a gaudy, abandoned mansion in Ohio. CNN's Karl Penhaul spoke to a shepherd who witnessed the final seconds of Germanwings Flight 9525. Penn Jillette was part of a panel speaking to CNN's Don Lemon about the controversial Indiana religious freedom law.

Intrinsic hallucinations: gaudy, abandoned

Extrinsic hallucination from TLHD-CNNDM.
Article: Gareth Barry has advised his Everton team-mate Ross Barkley against moving to Manchester City at this young stage of his career. Barry speaks from experience having spent four seasons at the Etihad before arriving on Merseyside and the veteran midfielder believes it is still too early for the 21-year-old to decide on his future. Ahead of the Toffees meeting with Manchester United on Sunday, Barry told the Mirror: 'Personally, I think he's still too young to make that move. Ross Barkley's rise to stardom has seen him repeatedly linked with Premier League champions Man City . Everton team-mate Gareth Barry has advised the youngster not to leave Goodison too soon . 'He's still learning the game. He's got the right manager here to push him to the next level. 'As soon as he reaches that next level, then there's another decision to be made. At the moment, I think it's too early.' And asked if considered the Premier League champions to be a graveyard for young talent, Barry added: 'I think so, yeah.' Barkley has overcome his early season struggles to play an influential role in Everton's recent revival and Barry believes the youngster he mentors daily can achieve anything he wants in the game. The 21-year-old signs autographs for fans after coming through a difficult start to the season . Veteran midfielder Barry spent four seasons at City before being found surplus to requirements . 'I sit next to him in the changing room at the training ground. I speak to Ross quite often,' said Barry. 'You feel sorry for him sometimes because the expectation is getting thrown on to his shoulders – people are expecting of him, week in, week out, goals and assists. 'That hasn't happened, but at the same time he's still improving as a player and growing in maturity. 'His ability and his strengths are there for everyone to see, he can go on and be a top top player.'
BART summary: Ross Barkley has been linked with a move to Manchester City. Everton team-mate Gareth Barry believes it is too early for the 21-year-old to leave Goodison Park. Barry spent four seasons at the Etihad before arriving on Merseyside. Everton face Manchester United at Old Trafford on Sunday.
Extrinsic hallucinations: at Old Trafford

C Human Annotation Details

Our human annotation was performed with 3 sets of 3 annotators, each annotating 50 examples. The full instructions are given below, together with an example of how the human annotation task looks.

Hallucination detection

This study evaluates hallucinations in automatic summarization models. A hallucination is information that is not directly supported by the article that the model has to summarize.
Main question: Can the summary sentence in question be inferred directly from the article?
There are two types of hallucinations: intrinsic and extrinsic hallucinations. They are defined as follows (from Maynez et al., 2020):
Intrinsic hallucination: Combination of information from the article that does not follow from it
Extrinsic hallucination: Information not present in the article
Not a hallucination: Paraphrases, or information directly inferred from the article
Importantly, this is not a question of whether the summary is true or false, just whether it faithfully represents the information in the article.

The goal in this study is to annotate a summary sentence with intrinsic and extrinsic hallucinations, by copying the words that cannot be inferred from reading the article. Here's an example (the part in red is the annotation that you will do [your annotations can stay black]):
Example annotation
Article: Manchester City was defeated by Crystal Palace 2-1 at the Etihad Stadium on Sunday. Glenn Murray and Jason Puncheon scored the goals for Palace, while Yaya Toure was the only scorer for City. City's best striker Sergio Aguero was left on the bench for yet another game. The result is especially shocking when comparing the squad's total transfer fees: £40m pounds for Crystal Palace vs. £500m for Manchester City.
Full summary: Crystal Palace beat Manchester City 2-1 on Saturday. Yaya Toure was left on the bench, and Crystal Palace have spent £40m on transfer fees so far this season.
Sentence in question: Yaya Toure was left on the bench, and Crystal Palace have spent £40m on transfer fees so far this season.
Intrinsic hallucinations: Yaya Toure
Extrinsic hallucinations: so far this season
Explanation: It was Sergio Aguero that was left

on the bench, not Yaya Toure (since he scored a goal, we know that he was playing). We're looking for a hallucination that is as small as possible, that's why we didn't mark "Yaya Toure was left on the bench", or "was left on the bench". For the extrinsic hallucination, there is no mentioning that the spending was for this season only. There is also a mistake in the first sentence of the summary (Saturday vs. Sunday in the article), but this is not the sentence in question, so we ignore it.

Notes

- If there are no hallucinations, leave the line blank.
- If there are multiple hallucinations in the sentence, separate them with a comma.
- Sometimes a sentence is not complete, or there are multiple sentences in one, but a period is missing to separate them. Just treat the "sentence in question" as if it were a single sentence. (These are artifacts of sentence splitting/the training data, which we do not evaluate here.)
- The examples below have a visual help: Text overlaps of more than two words between the sentence and the article are written in **bold** and numbered at the end, like this: [1]. This is just a help for you to find information faster, and does not mean the model copied the parts from there. Example: **Article: This year's harvest was**[1] especially rich **on apples.**[2] **Sentence: This year's harvest was**[1] high **on apples.**[2]
- Hint: Read the sentence in question first, and then look for the relevant information in the article.