# Think While You Write: Hypothesis Verification Promotes Faithful Knowledge-to-Text Generation

[1]Yifu Qiu, [2]Varun Embar, [1] Shay B. Cohen, [2]Benjamin Han

[1]Institute for Language, Cognition and Computation, University of Edinburgh

[2]Apple Inc.

## Highlights

- Knowledge-to-text generators often struggle to faithfully generate descriptions for the input facts: they may produce hallucinations that contradict the input, or describe facts not present in the input. We propose **TWEAK decoding**, which can be integrated with any generator without retraining.
- TWEAK treats the generated sequences at each decoding step and its future sequences as *hypotheses*, and ranks each generation candidate based on the extent to which their hypotheses are supported by the input facts using a Hypothesis Verification Model (HVM).
- We test TWEAK with two generators, and the best TWEAK variants improve on average for the two models by 2.24/7.17 points in faithfulness (FactKB) in in/out-of-distribution evaluations, respectively, and with only a 0.14/0.32-point decline in quality (BERTScore)

## Motivation

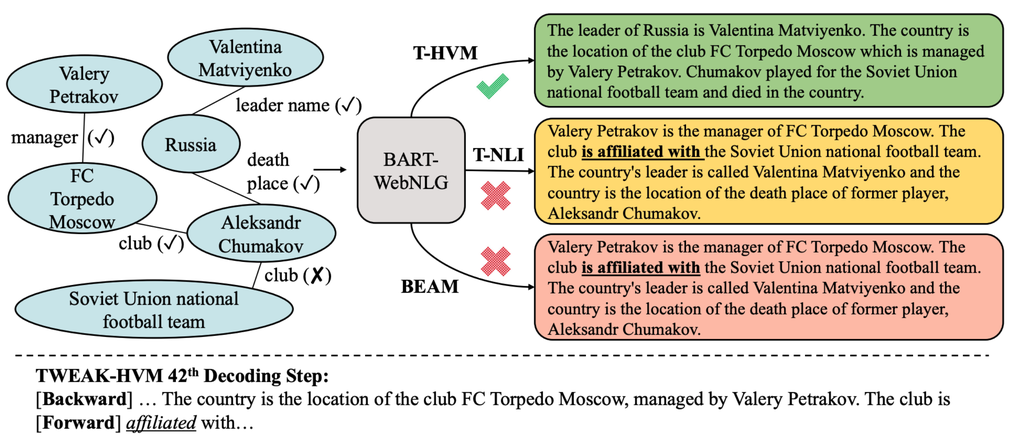



Figure 1. Qualitative examples for 3 decoding methods.

- On the left are the input facts, and on the right are the natural language descriptions generated from three approaches: T-HVM (TWEAK-HVM), T-NLI and baseline Beam Search (the first two are TWEAK variants). Both T-NLI and Beam Search generated "affiliated with…" due to hallucination from the fact (Aleksandr Chumakov, club, Soviet Union national football team) (the triple is marked with X).

## TWEAK Decoding Framework

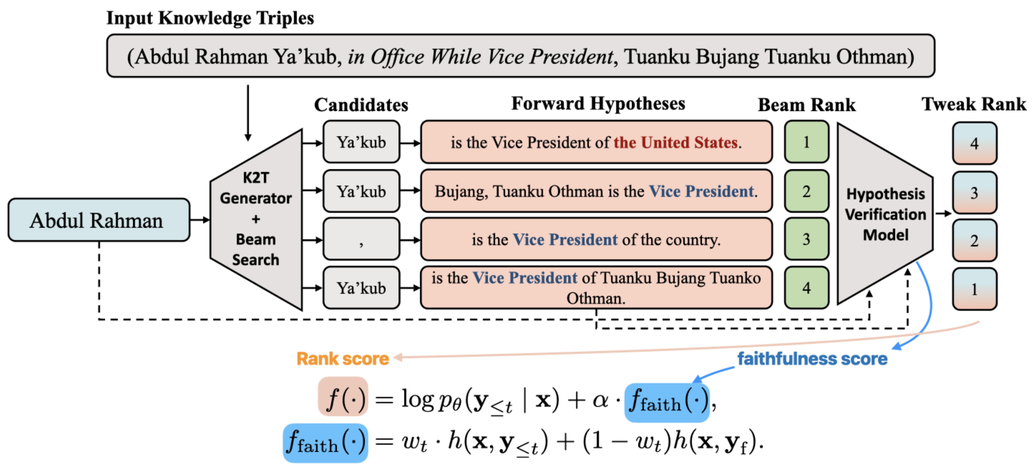


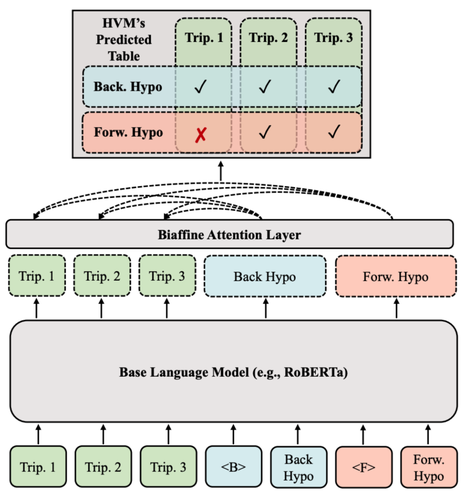

Figure 2. Illustration of TWEAK decoding framework.

- Instead of solely ranking candidates based on generative model's predicted likelihood, TWEAK ranks them by considering faithfulness, which is estimated by assessing the backward and forward hypotheses of each candidate with a Hypothesis Verification Model (HVM).
- In the 4th decoding step of this example, the beam search promotes the candidate leading to hallucinations (e.g., "United States"), but not TWEAK.

## Task-specific HVM

| FATE Instance | Type | Label |
|---|---|---|
| *(Ireland, largest_city, Dublin)* | PTs | - |
| *(Ireland, national_capital, Dublin)* | NTs | - |
| *Dublin is Ireland's <S0> largest city </S0>* | PD | - |
| *Dublin is Ireland's <S0> national capital </S0>* | ND | - |
| **Synthesized Hypotheses (at $10^{th}$ Decoding Step)** | | |
| *Dublin is Ireland's largest largest city.* | BH<br>FH | ✓<br>✓ |
| *Dublin is Ireland's national national capital.* | BH<br>FH | ✗<br>✗ |

Although an off-the-shelf Natural Language Inference (NLI) model can be used as the HVM, we train a more efficient and more fine-grained task-specific HVM using our novel dataset, FATE (Fact-Aligned Textual Entailment). Hallucinated spans are marked in FATE, and HVM is trained to predict whether a hallucination occurs word-by-word. The model is more efficient because facts and hypotheses are all encoded and classified in one go.




## Paper & Code

(a) Paper

(b) Github Repo

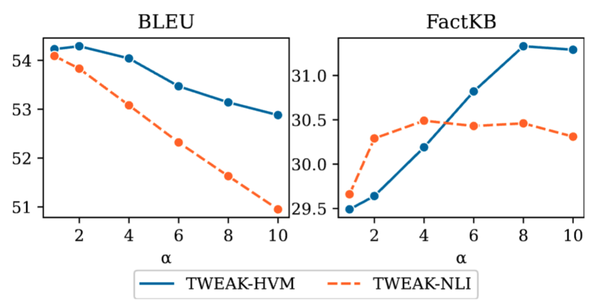
## TWEAK Enhances Faithfulness

| | Decoding | FactKB | BLEU | MET | BertScore |
|---|---|---|---|---|---|
| BART-large | Greedy | 27.74 | 51.3 | 66.79 | 94.2 |
| | Beam | **28.91** | **54.23** | **67.55** | **94.35** |
| | TWEAK-NLI-F | 30.46 | 52.02 | 67.17 | 94.2 |
| | TWEAK-NLI-B | 30.59 | 49.68 | 65.88 | 94.12 |
| | TWEAK-NLI-B+F | 30.47 | 51.62 | 66.84 | 94.19 |
| | TWEAK-HVM | **31.34** | **53.14** | **67.38** | **94.25** |
| T5-large | Greedy | 30.14 | 57.71 | 68.71 | 94.84 |
| | Beam | **31.29** | **58.93** | **69.38** | **94.86** |
| | TWEAK-NLI-F | 33.03 | 53.51 | 67.8 | 94.39 |
| | TWEAK-NLI-B | 31.49 | 44.96 | 65.02 | 93.93 |
| | TWEAK-NLI-B+F | 32.71 | 51.71 | 66.73 | 94.19 |
| | TWEAK-HVM | **33.34** | **57.31** | **69.02** | **94.68** |

Table 1. Results of decoding baselines and our TWEAK decoding variants measured by faithfulness metric (**FKB** = FactKB) and quality metrics (**BLEU**, **MET** = METEOR, **BS** = BERTScore) on WebNLG. Numbers in **bold** are the highest scores among the baselines or among the TWEAK variants.

- In in-distribution setting TWEAK outperforms the baseline on faithfulness by +2.22 points while lowering only 0.14 points on quality.

## Analysis




| #Triples | | Short | Medium | Long |
|---|---|---|---|---|
| #Sample | | 908 | 2196 | 620 |
| BLEU | HVM | 64.18 | 50.22 | 46.26 |
| | NLI | 63.48 | 48.59 | 45.23 |
| | Δ | +1.09% | +3.25% | +2.23% |
| FactKB | HVM | 18.11 | 33.47 | 43.17 |
| | NLI | 18.06 | 32.67 | 40.81 |
| | Δ | +0.28% | +2.39% | +5.47% |

Figure 4. **Left**: The effect on quality (BLEU) and faithfulness (FactKB) from choosing different weighting coefficient. **Right**: TWEAK decoding performance on WebNLG with increasing number of input triples. We split the WebNLG into three groups: Short (1 triples), Medium (2-4 triples) and Long (5-7 triples).

**Weighting Effects.** We observe that increasing the weight on faithfulness score improves faithfulness in almost all settings at the cost of reduced quality.

**TWEAK is Effective on More Fact Triples.** On generative quality (BLEU) we observe that TWEAK-HVM outperforms TWEAK-NLI-B+F by a similar amount across the three groups. On faithfulness (FactKB), however, TWEAK-HVM's improvement over TWEAK-NLI-B+F is positively correlated with the number of input triples.

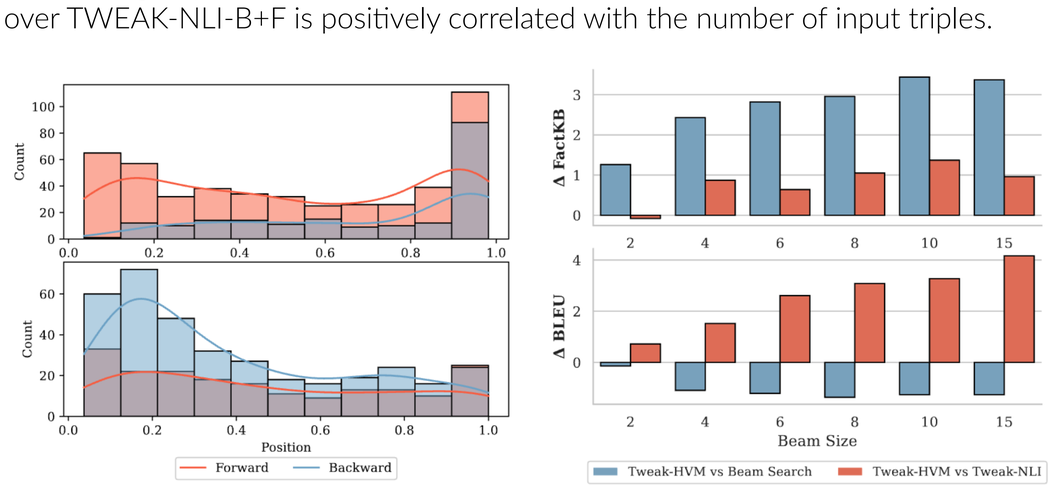



Figure 5. **Left**: The distributions of the relative positions where negative predictions (i.e., possible hallucination) happen during the decoding process. 0 and 1 along the horizontal axis represent the start and end of the decoding. The upper and bottom panel represent TWEAK-HVM and TWEAK-NLI-B+F running on WebNLG with BART-large, respectively. **Right**: Performance differences (Δ) on quality (BLEU) and faithfulness (FactKB) between TWEAK-HVM, TWEAK-NLI-B+F and beam search on various beam sizes $\{2, 4, 6, 8, 10, 15\}$.

**Where are hallucinations found?** TWEAK-HVM predicts more hallucinating forward hypotheses, while TWEAK-NLI-B+F leans towards more hallucinating backward hypotheses. This divergence can be attributed to the training differences between NLI and HVM.

**TWEAK with Larger Beam Size.** TWEAK-HVM has a greater capacity in taking advantage of a bigger beam size.