A Comprehensive Study of Hallucinations in Neural Machine Translation

Nội dung

1. Taxonomy of Translation Pathologies
2. Hallucination Detection Methods
3. Hallucination datasets
4. DeHallucinator: Overwriting Hallucinations at Test Time

Taxonomy of Translation Pathologies

Hallucination

Oscillatory: Lặp lại bất thường các từ và cụm từ.

- Ex: “I love you” -> “Tôi tôi yêu yêu bạn”

Largely fluent: Bản dịch trôi chảy nhưng lại không liên quan đến câu nguồn (strongly / fully)

- Ex: “Happy birthday” -> “Chúc mừng năm mới” / “Bạn có khỏe không”

Taxonomy of Translation Pathologies

Translation errors

- **Undergeneration:** Dịch không đầy đủ, không bao quát nội dung câu nguồn nhưng nội dung còn lại vẫn còn liên quan đến câu nguồn => không phải ảo giác.
- **Mistranslation of named entities:** Không được coi là ảo giác bởi vì nó không tách biệt so với câu nguồn mà chỉ là dịch sai hoặc thiếu 1 phần nhỏ (tên riêng, địa chỉ,…)

Hallucination Detection Methods

Quality Filters

- **Reference-free methods:** Sử dụng mô hình **COMET-QE** để đánh giá chất lượng mà không cần bản dịch tham chiếu.
- **Reference-based methods:** Cần bản dịch tham chiếu (COMET, CHRF2, …)

Hallucination Detection Methods

Heuristics

- **Binary-score Heuristics**
  - **Top n-gram count (TNG)**
  - **Repeated targets (RT)**

Top n-gram count (TNG)

Trong bản dịch có cụm từ lặp đi lặp lại **nhiều hơn mức bất thường** so với cụm lặp trong câu gốc, thì bản dịch có khả năng là **hallucination.**

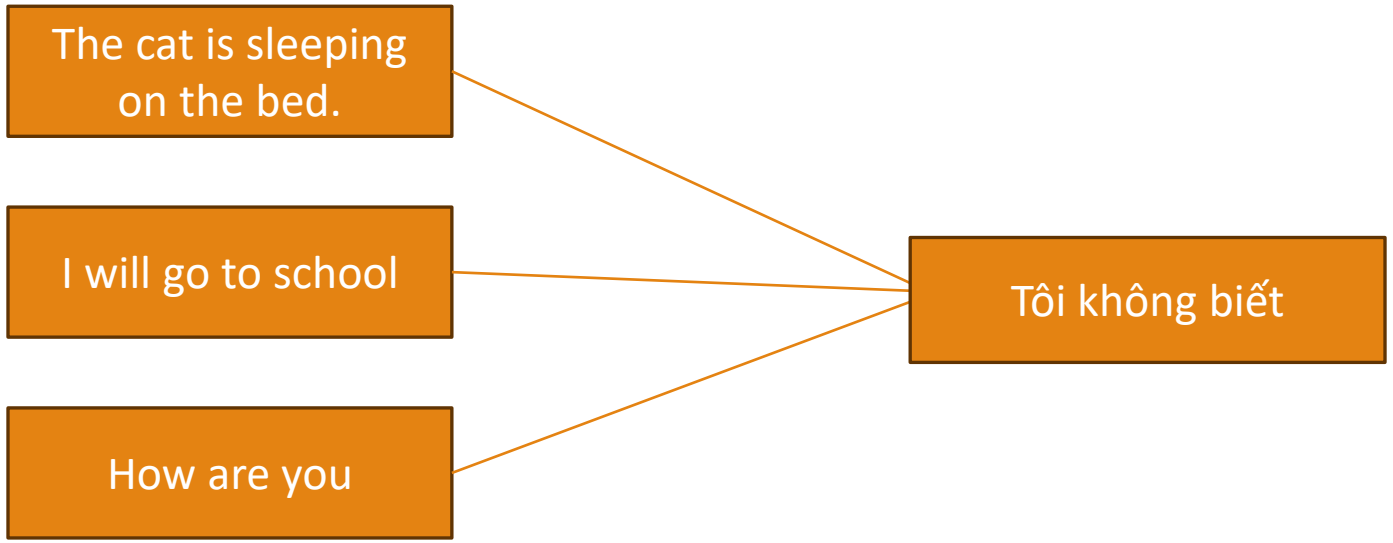
Ví dụ:

- Source (EN): The economic situation is improving in many countries around the world.

- Translation (VI): Tình hình kinh tế đang cải thiện trên **toàn thế giới**, **toàn thế giới, toàn thế giới, toàn thế giới**.

Repeated targets (RT)

Nhiều câu nguồn chi được dịch thành đúng 1 câu đích.



Hallucination Detection Methods

Heuristics

- **Anomalous decoder-encoder attention**
  - **Attn-to-EOS:** tỷ lệ attention của decoder tập trung vào token **EOS** trong câu nguồn
  - **Attn-ign-SRC:** tỷ lệ từ trong source có **tổng (attention < threshold)**

Hallucination Detection Methods

Uncertainty-Based Heuristics

Sequence log-probability (Seq-Logprob):

- Halluc -> model không tin tưởng -> logprob thấp

$$\frac{1}{L}\sum_{k=1}^{L}\log P(y_k \mid y_{<k}, x, \theta).$$

Dissimilarity of MC hypotheses (MC-DSim):

- Sinh ra nhiều bản dịch bằng MC-dropout sau đó tính giá trị trung bình SIM (METEOR, BERTScore)

$$\frac{1}{N}\sum_{i=1}^{N}\mathrm{SIM}(h_i, y).$$

Ví dụ

- **Oscillatory**

```
ID: 9314, En: NEVER CALL ME AGAIN !
 Vi: ĐỪNG BAO GIỜ GỌI CHO TÔI NỮA !
 Trans: ĐƯỜNG THƯỜNG TÂN ĐƯỜNG TÂN ĐƯỜNG TÂN ĐƯỜNG TÂN ĐƯỜNG TÂN ĐƯỜNG TÂN !
 seq_logprob: -2.404375
```

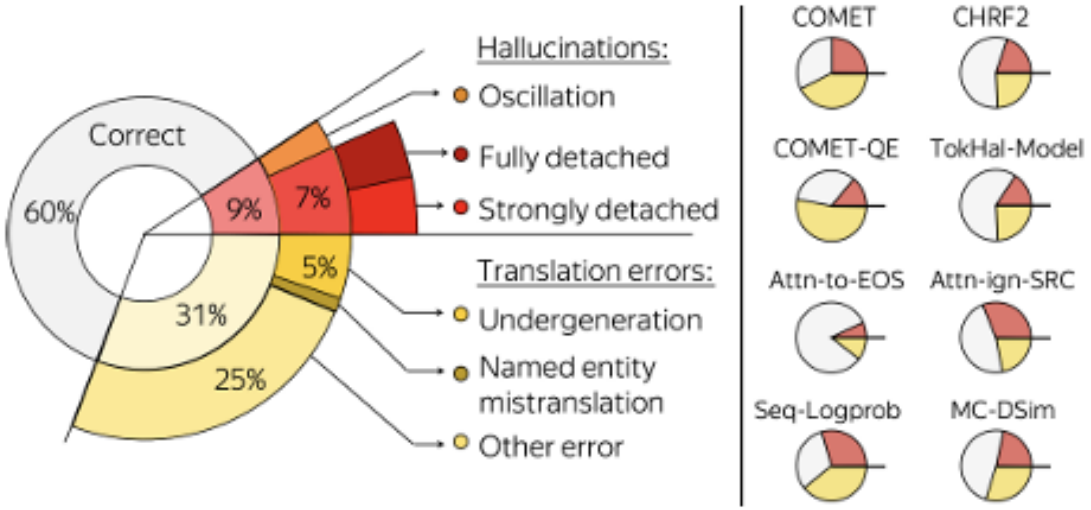
- **Strongly detached**

```
ID: 18451, En: THE MAC IS BORN
 Vi: MÁY TÍNH MAC XUẤT HIỆN
 Trans: CHƯỜNG TRƯỜNG TÂN
 seq_logprob: -2.3167
```

- **Fully detached**

```
ID: 4986, En: How was that for poignant ?
 Vi: Như thế sao có thể gọi là mỉa mai ?
 Trans: Làm thế nào để đánh bom được ?
 seq_logprob: -1.7180666666666662
```

Hallucination datasets



-> Seq-Logprob cho kết quả tốt nhất

- **Tìm kiếm nhiều bản dịch có khả năng hallucination.**
  - Heuristics
  - Quality filter
  - TNG / RT
- **Chọn một số bản dịch từ "long tail"** tức là bản dịch có điểm rất thấp trong phân phối điểm dự đoán của các hệ thống phát hiện lỗi.
- **Gán nhãn dữ liệu thủ công.**

DeHallucinator: Overwriting Hallucinations at Test Time

Các bước thực hiện:

- Phát hiện hallucination bằng seq-logprob

- Tạo ra nhiều bản dịch bằng MC-Dropout

- Chấm điểm các bản dịch bằng COMET-QE / Seq-Logprob

- Thay thể bản cũ bằng bản dịch mới tốt hơn