Title

VisualGround: A Multimodal Framework for Evaluating Machine Translation with Visual Consistency

Problem Statement

Current machine translation evaluation metrics struggle to detect and penalize hallucinations, especially when the translated content appears fluent but deviates semantically from the source. This is particularly problematic for content with strong visual components, where semantic consistency is crucial.

Motivation

Existing methods primarily rely on text-based comparisons between source, reference, and candidate translations, which may miss subtle semantic divergences. Incorporating visual information can provide an additional modality to ground translations and detect semantic inconsistencies that may not be apparent from text alone. By leveraging the rich semantic information in images, we can create a more robust evaluation framework that better aligns with human judgments, especially in detecting and penalizing hallucinations.

Proposed Method

We propose VisualGround, a multimodal translation evaluation framework that leverages both textual and visual information. The method consists of the following key components: 1) Dataset Creation: We create a large-scale dataset of image-caption pairs in both source and target languages. 2) Cross-lingual Vision-Language Model: We train a model that can encode both text and images across languages. 3) Visual Representation Comparison: During evaluation, we use the trained model to generate visual representations for the source text, reference translation, and candidate translation. We compare these visual representations to detect semantic divergences. 4) Image Retrieval: We use the trained model to retrieve the most relevant images for each text, comparing the retrieved images across source, reference, and candidate to identify potential hallucinations. 5) Scoring: We combine the visual representation comparison and image retrieval results to produce a final score that reflects the translation quality and semantic consistency.

Step-by-Step Experiment Plan

Step 1: Dataset Creation

- 1 Collect a large-scale dataset of image-caption pairs in English and the target language (e.g., German). Use existing datasets like Multi30K, MSCOCO, and Flickr30K as starting points. 1.2 Use professional translators to create high-quality translations for captions that don't have corresponding translations. 1.3 Ensure diversity in the dataset by including various domains and visual concepts. 1.4 Split the dataset into train, validation, and test sets.

Step 2: Cross-lingual Vision-Language Model Training

- 1 Choose a pre-trained vision-language model (e.g., CLIP) as the starting point. 2.2 Fine-tune the model on our created dataset to enable cross-lingual capabilities. Use contrastive learning objectives to align visual and textual representations across languages. 2.3 Implement data augmentation techniques such as random cropping, color jittering for images, and back-translation for text to improve robustness. 2.4 Monitor training progress using validation set performance on image-text retrieval tasks in both languages.

Step 3: Visual Representation Comparison

- 1 For a given translation triple (source, reference, candidate), use the trained model to generate visual representations. 3.2 Implement multiple similarity metrics (e.g., cosine similarity, Euclidean distance) to compare the visual representations. 3.3 Focus on detecting discrepancies in entities, actions, and attributes by analyzing attention maps or using probing tasks. 3.4 Develop a scoring mechanism that quantifies the degree of semantic divergence based on these comparisons.

Step 4: Image Retrieval

- 1 For each text (source, reference, candidate), use the trained model to retrieve the top-k most relevant images from a large image database. 4.2 Implement metrics to compare the retrieved image sets (e.g., Jaccard similarity, rank correlation). 4.3 Develop a scoring mechanism that quantifies the consistency of retrieved images across the translation triple.

Step 5: Scoring Integration

- 1 Design a composite score that combines the visual representation comparison and image retrieval results. 5.2 Implement different weighting schemes for the components and experiment with their impact on the final score. 5.3 Calibrate the scoring mechanism using a held-out validation set with human judgments.

Step 6: Evaluation

- 1 Apply VisualGround to existing machine translation datasets augmented with relevant images (e.g., Multi30K). 6.2 Compare VisualGround against text-only baselines like BLEU, METEOR, and COMET, as well as reference-free metrics. 6.3 Conduct a human evaluation study to assess the correlation of VisualGround with human judgments, particularly focusing on hallucination detection accuracy. 6.4 Analyze the performance of VisualGround across different types of translations, including those with subtle semantic errors that might be missed by text-only metrics.

Step 7: Ablation Studies and Analysis

- 1 Conduct ablation studies to quantify the contribution of each component (visual representation comparison, image retrieval) to the final score. 7.2 Analyze the impact of the cross-lingual vision-language model's architecture and training data on the evaluation performance. 7.3 Investigate how VisualGround performs on different types of visual content (e.g., concrete objects vs. abstract concepts) and translation errors.

Test Case Examples

Baseline Prompt Input (BLEU Score)

Source: A red car is parked on the street. Reference: Ein rotes Auto parkt auf der Straße. Candidate: Ein blaues Auto fährt auf der Autobahn.

Baseline Prompt Expected Output (BLEU Score)

BLEU Score: 0.223

Proposed Prompt Input (VisualGround)

Source: A red car is parked on the street. Reference: Ein rotes Auto parkt auf der Straße. Candidate: Ein blaues Auto fährt auf der Autobahn.

Proposed Prompt Expected Output (VisualGround)

VisualGround Score: 0.312 Visual Representation Similarity: 0.456 Image Retrieval Consistency: 0.168 Detected Inconsistencies: Color (red vs. blue), Action (parked vs. driving), Location (street vs. highway)

explanation

While BLEU gives a moderate score due to some word overlap, VisualGround detects the semantic divergences in color, action, and location by leveraging visual information. The low Image Retrieval Consistency score indicates that the images retrieved for the candidate translation are significantly different from those for the source and reference, suggesting a potential hallucination.

Fallback Plan

If VisualGround doesn't significantly outperform existing metrics, we can pivot to an analysis paper that provides insights into the challenges of incorporating visual information in translation evaluation. We could investigate cases where visual grounding succeeds or fails, analyze the types of translation errors that are most amenable to visual detection, and explore the limitations of current vision-language models in this context. Additionally, we could expand our study to include more languages and diverse visual domains to understand how the effectiveness of visual grounding varies across different linguistic and cultural contexts. This analysis could provide valuable insights for future research in multimodal machine translation evaluation.

Ranking Score: 6