Интерпретация результатов дообучения языковых моделей (Трансформеров)

Interpreting the Fine-Tuning Results of Language Models (Transformers)

Туеста Льаха Хуан Карлос
tuesta.lx@phystech.edu



Introduction

- **Objective**: Development of a method for constructing an interpretable description of the state of a Transformer network and identifying causal relationships between the states and the network's performance outcomes.
- **Focus**:
  - Introduction to paraphrase generation.
  - Main approaches and methods.
  - Interpretability in language models.
  - Evaluation metrics for accuracy.

Language Models and Their Application in NLP

- Language models predict the probability of word sequences.
- The Transformer architecture has significantly improved natural language processing.
- Key example: BERT and its adaptability to different languages.
- **Relevant Study**: "What the [MASK]? Making Sense of Language-Specific BERT Models"(Nozza et al., 2020) on BERT and monolingual models.

Techniques for Paraphrase Generation

- Paraphrase generation is essential for tasks like automatic rewriting and data augmentation.
- **AESOP Approach**: Adaptive syntactic control for generating coherent paraphrases.
- **Technique by Liu et al. (2023)**: Unsupervised paraphrase generation using syntactic knowledge.

AESOP

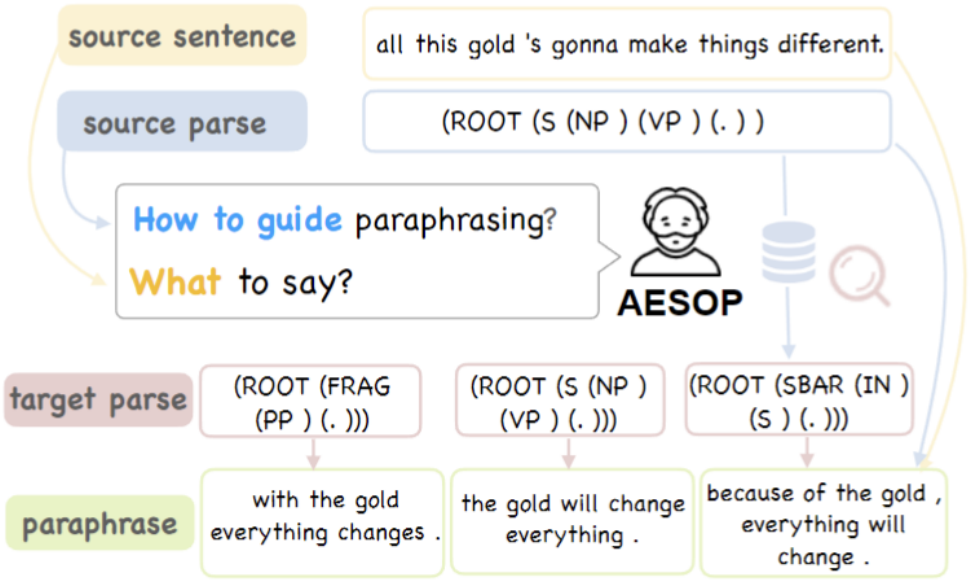

Рис.: Given a source sentence, AESOP selects tar- get syntactic parses adaptively to guide paraphrase generation. Paraphrases here are all generated by AESOP, which preserve the semantics from source sentences and conform with the selected syntactic parses

Unsupervised paraphrase

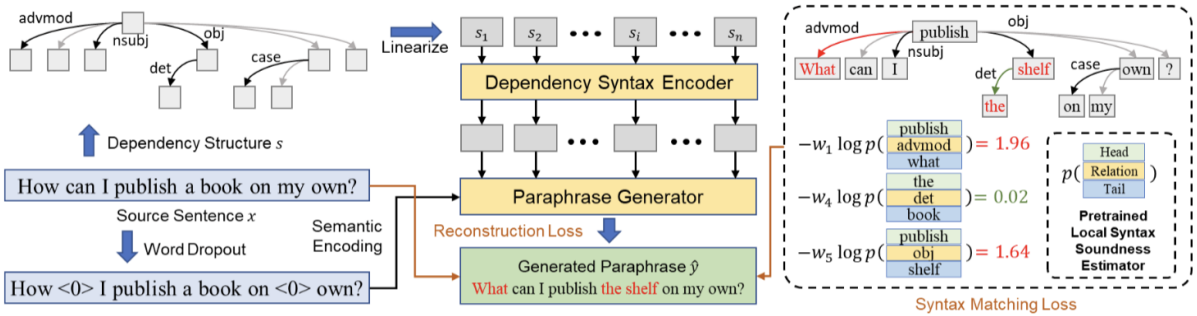

Рис.: This model generates paraphrases without relying on labeled data. It leverages syntactic information to maintain grammatical correctness and semantic similarity.

Explainability of Large Language Models (LLMs) and Accuracy Evaluation

- Explainability is crucial for trusting language models.
- Interpretability methods include attention visualization and saliency methods.
- Importance of comparing models with metrics like BLEU and ROUGE.
- **Key Study**: "Explainability for Large Language Models: A Survey"(Zhao et al., 2024) on explainability in LLMs.

Structural Probing and Syntax in Word Representations

- Multilingual models like BERT learn common linguistic representations.
- Monolingual models capture better syntactic dependencies specific to a language.
- **Key Study**: "A Structural Probe for Finding Syntax in Word Representations"(Hewitt and Manning, 2019) on structural probes.

Challenges and Opportunities in Paraphrase Generation

- Challenges: Maintaining original meaning without introducing ambiguity.
- Generating grammatically correct and fluent paraphrases remains a challenge.
- Opportunity: Improvement in unsupervised models like the approach by Liu et al. (2023).

Conclusion of Theoretical Framework

- Transformer-based LLMs have significantly advanced NLP.
- Paraphrase generation and explainability are crucial to improving model reliability.
- Further research is needed on syntactic representations and adaptive control in paraphrase generation.

Спасибо за внимание!