Self-Ensemble of N-best Generation Hypotheses by Lexically Constrained Decoding

Ryota Miyano†, Tomoyuki Kajiwara‡, Yuki Arase†

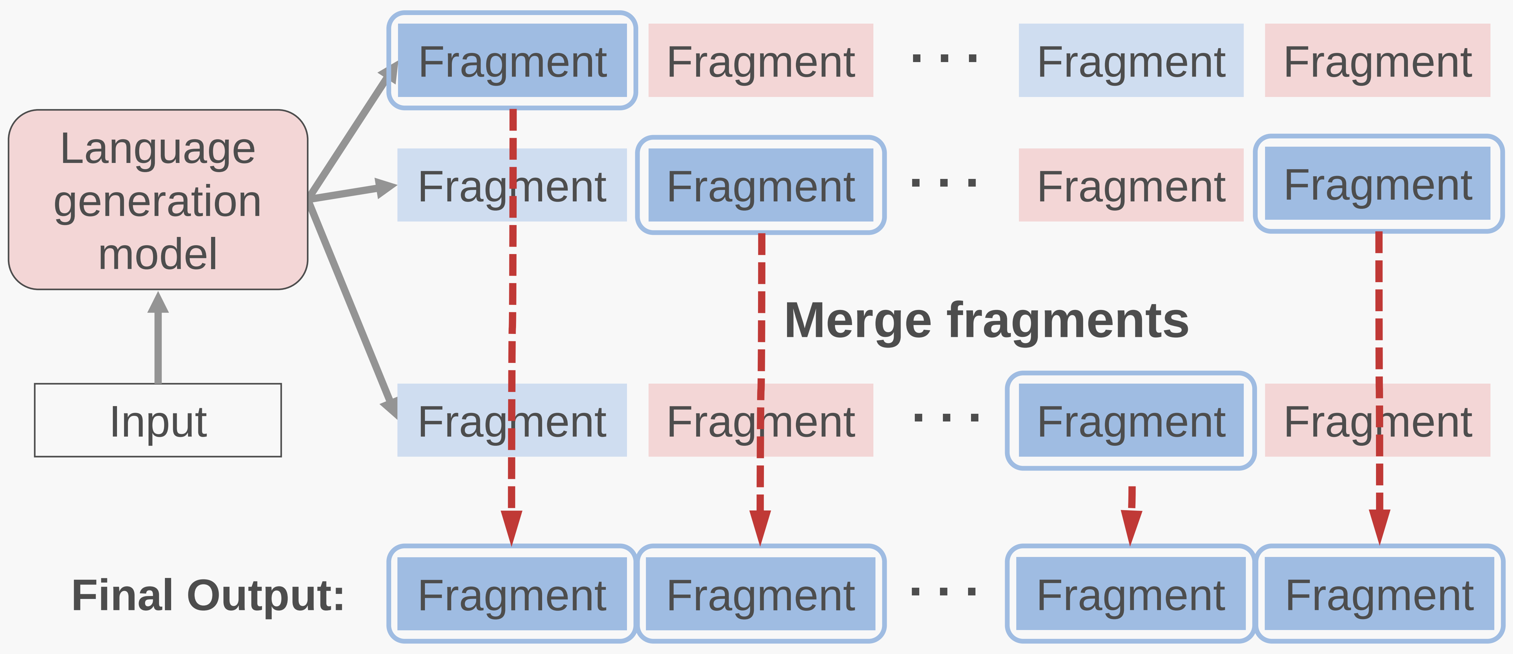
†Graduate School of Information Science and Technology, Osaka University
‡Graduate School of Science and Engineering, Ehime University

Code, Paper

Introduction

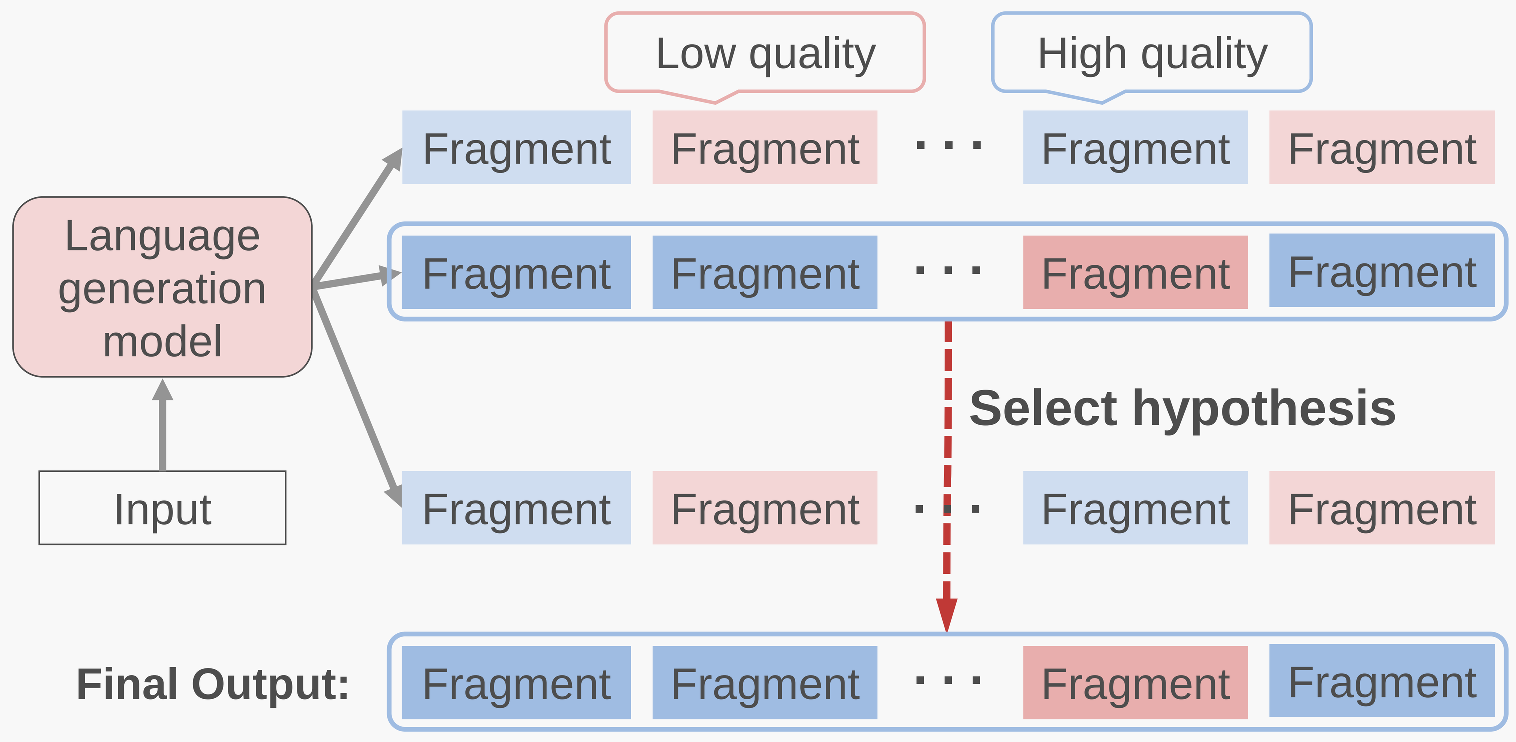
Our approach

- Assumption: There exists ***partly*** higher quality hypotheses
- **Merge high quality fragments** to obtain better output
- Use **lexical constraints** to control output



Existing method – Reranking

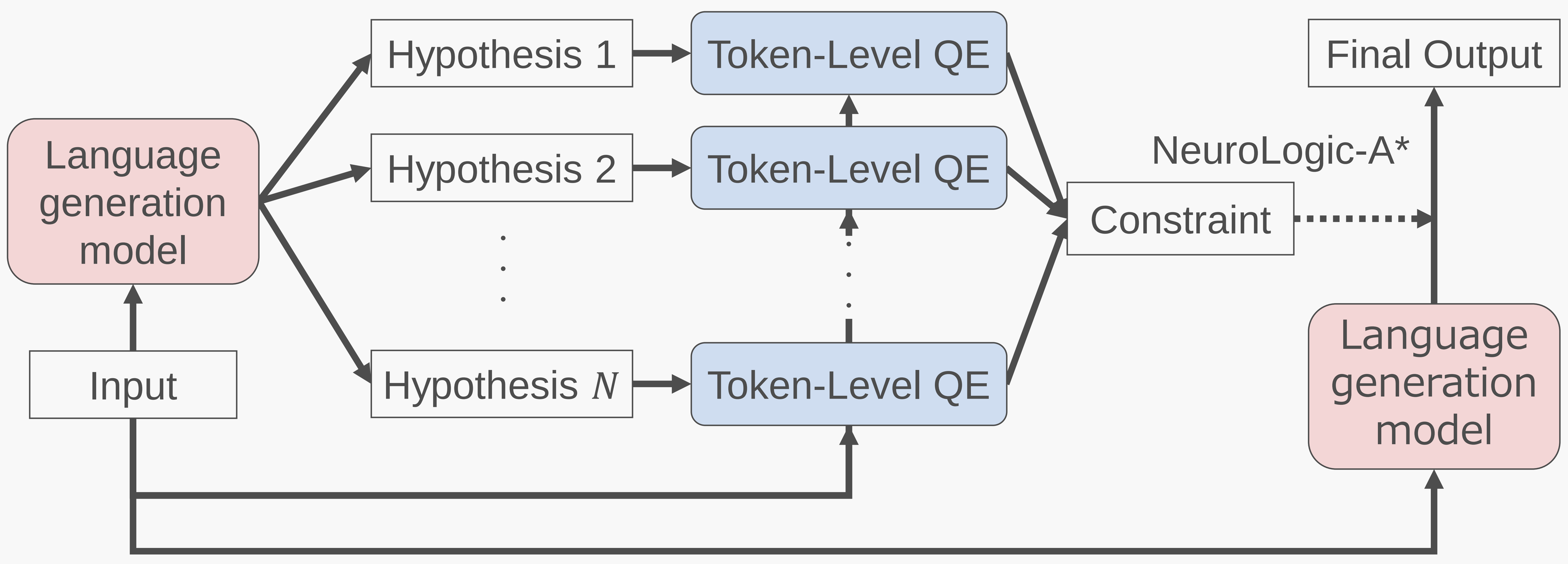
- Assumption: There exists a higher quality hypothesis
- Re-evaluate hypotheses and **select the best hypothesis**



Proposed method

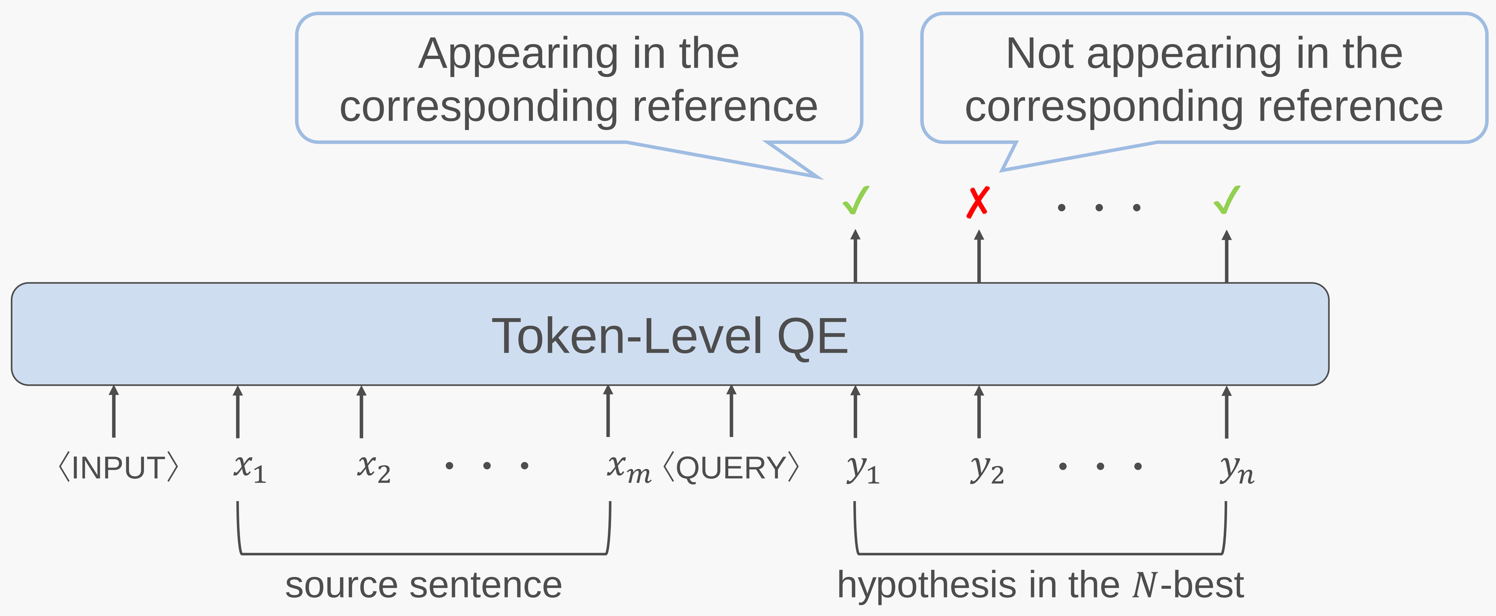
Overview

- Create constraints based on the **token-level QE model**
  - QE model predict whether each token in N-best hypotheses should be used or avoided in the final output
- Generate final output using Lexically constrained decoding
  - NeuroLogic-A* (Lu et al., 2022)



Training of the token-level QE model

- Create training data
  - Generate N-best hypotheses for training sentences using a language generation model
  - Assign positive labels to tokens appearing in the reference and negative labels to tokens not appearing
- Fine-tune a pretrained masked language model

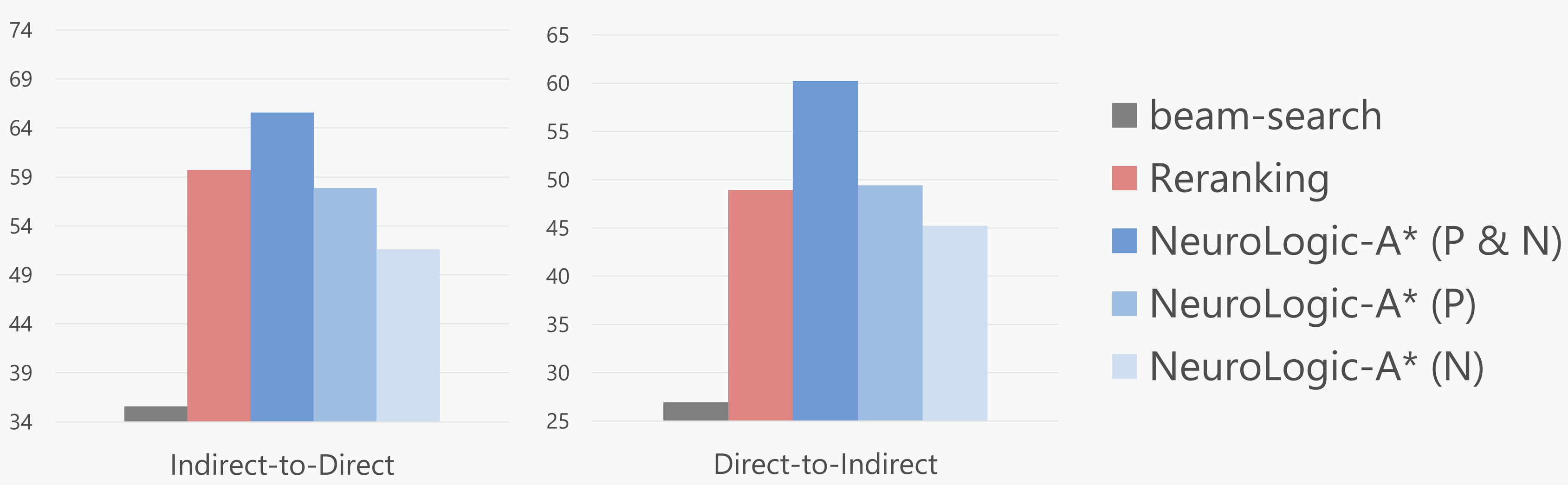


Experiments

Evaluate our method on language generation tasks

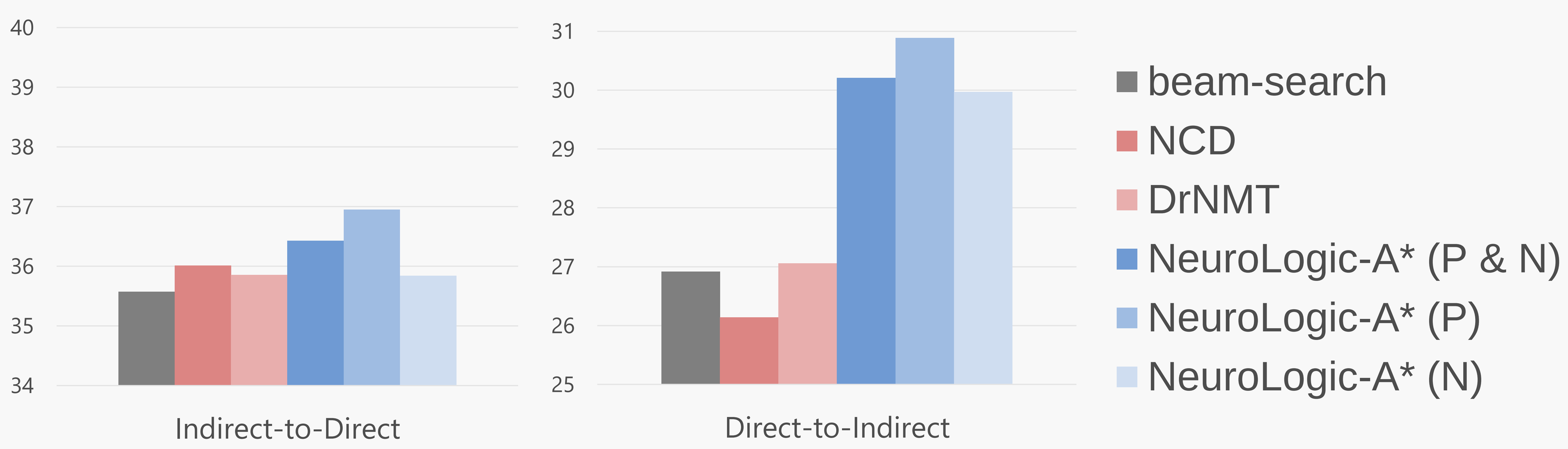
- Paraphrasing
  - Dataset: DIRECT (Takayama et al., 2021)
  - Metrics: BLEU
- Summarisation
  - Dataset: XSum (Narayan et al., 2018)
  - Metrics: ROUGE-L

Paraphrasing - Oracle (Assume the reranker and QE model are ideal)

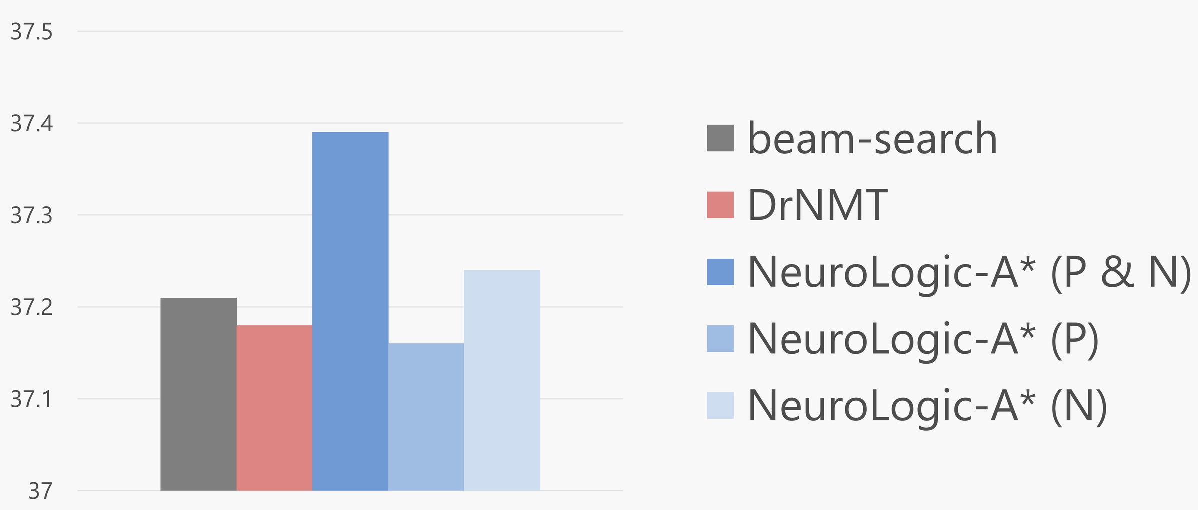


Ensembling N-best hypotheses is more effective than selecting the best hypothesis

Paraphrasing



Summarisation



Proposed method significantly outperforms the strong reranking methods