Levenshtein Training for Word-level Quality Estimation

Shuoyang Ding  Marcin Junczys-Dowmunt
Matt Post    Philipp Koehn

EMNLP 2021
Punta Cana, Dominican Republic
Problem Definition

- Word-level Quality Estimation (QE)
  - Generate **binary quality tags** for each **word** and the **gaps** between every two words
  - **BAD** on **word** indicates deletion/substitution
  - **BAD** on **gap** indicates insertion

http://statmt.org/wmt21/quality-estimation-task.html
Evaluation

- Ask **humans** to **post-edit** system translation
- Generate word-level references by computing **TER**
Predictor-Estimator

Picture from Lee (2020)
Translation Probability?

- Problems:
  - **Left-context only**, while bidirectional available
  - Cannot predict **gap labels**
Can we do better?

- **Levenshtein Transformer!**

- A non-autoregressive translation model that performs decoding iteratively

Gu, Wang and Zhao (2019)
Why LevT?

- Solves both weaknesses:
  - **Mask prediction** -> can predict **gap labels**
  - **Iterative decoding** -> access to **bidirectional** context
Training Process

(Optional) Stage 0
Pre-train Encoder/Decoder

Stage 1
Levenshtein Pre-training

Stage 2
Pseudo Translation Triplet Fine-tune

Stage 3
Human Translation Triplet Fine-tune
Training Process

- Training a LevT for translation

- Knowledge distillation (KD) is often necessary

- (Optionally) Initialize the transformer blocks of encoder/decoder with pretrained model
What’s Left?

- Iterative decoder:
  - Noised target -> Subword-level Edit Tags

- QE decoder:
  - MT target -> Word-level Edit Tags

- Not great. Should minimize this mismatch.
Training Process

- Fine-tune with **pseudo translation triplets**
  (src, tgt, pe)
Training Process

- Fine-tune with the actual human post-edited translation triplet
- (7k triplets)
Experiments

• We evaluate under the WMT20 shared task setup
  • **En-De** and **En-Zh** language pairs
• Metric: Matthews Correlation Coefficient (**MCC**)  
  • **higher is better**
Constrained Setting

![Graph showing MCC scores for different models in en-de and en-zh settings.](image)

- **OpenKiwi**
- **Levenshtein w/o KD**
- **Levenshtein**
- **OpenKiwi**
- **Levenshtein**

**Legend:**
- **MCC**
- **en-de**
- **en-zh**
Unconstrained Setting

Ding et al. (2021) – Levenshtein Training for Word-level Quality Estimation
Conclusion

- We propose to use **Levenshtein Transformer** for the task of **word-level QE**

- We propose a series of **fine-tuning procedures** to adapt the LevT model to match the input/output format of the word-level QE

- **Data-efficient**

- **Competitive result**
WMT21 QE Shared Task

- LevT-QE participated in **WMT21 QE shared task**

- Results are competitive, achieving the **1st place** in Word-MCC metric for en-de

- System paper: *The JHU-Microsoft Submission for WMT21 Quality Estimation Shared Task*

- **Thursday, Nov 11th, 10:30a – 12:00p AST**
Thanks!
Jobs@NYC?

email: dings@jhu.edu
twitter: @_sding
github: shuoyangd
website: sding.org

paper & code