Edit Distance Based Curriculum Learning for Paraphrase Generation

Sora Kadotani* Tomoyuki Kajiwara** Yuki Arase* Makoto Onizuka*

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

Paraphrase Generation

- Can be formulated as a monolingual machine translation problem
- Application examples
  - English education support
  - Preprocessing for machine translation

I looooove this song SOOO Much!!!!!!

Informal → Formal
Formality Transfer

I love this song very much.

Curriculum Learning

- Training a model starting from easy samples and gradually moving to difficult ones
- Curriculum learning for machine translation [1]
  - Improve translation quality
  - The metric of difficulty
    - Sentence Length (SL)
    - Word Rarity (WR)

Easy	Medium	Difficult
Thank you.	Thank you very much.	Thank you for your helping me with my work.

Training Time

Difficulty of Paraphrase Generation

- Paraphrases require a few transformations:
  Copy almost all the input sentence's words (**Easy**)
- Paraphrases require drastic transformations:
  Require complex rewriting operations (**Difficult**)
- We propose to estimate the difficulty of transformations by **edit distance**

Source Sentence	Target Stentence
Their first two albums were **pretty** good.	Their first two albums were **very** good.
no where there is no such thing	That does not exitst.

Edit Distance

- Number of edit operations required to convert the word sequence
- Small edit distance: **Easy**
- Large edit distance: **Difficult**

Edit distance = 1

Source Sentence	Target Stentence
Their first two albums were **pretty** good.	Their first two albums were **very** good.
no where there is no such thing	That does not exitst.

Edit distance = 7

Proposed Method

Edit Distance Based Curriculum Learning

- Apply curriculum learning to paraphrase generation
- Use the existing curriculum learning framework [1]
- The metric of difficulty
  - Edit Distance (ED)

Experiment on Formality Transfer

Evaluate the quality of paraphrase generation

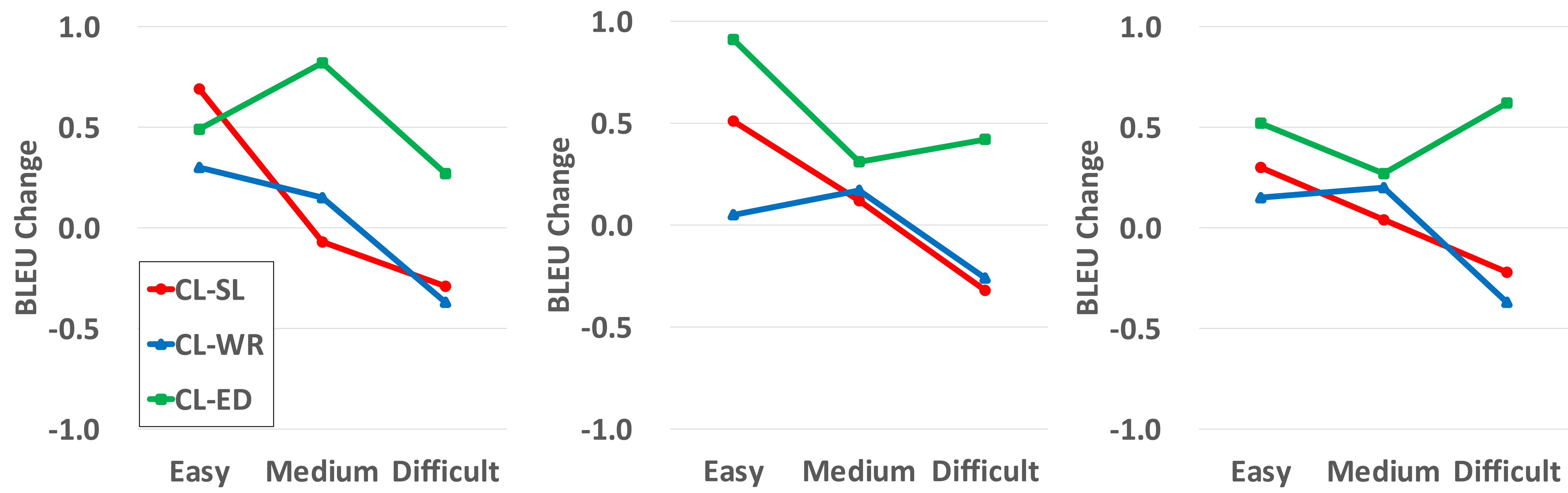
- Setup
  - Dataset: GYAFC [2]
  - Model: Transformer
  - Evaluation metric: BLEU
- Comparison methods
  - Baseline: Without curriculum learning
  - CL-SL: Curriculum learning with sentence length
  - CL-WR : Curriculum learning with word rarity
  - CL-ED : Curriculum learning with edit distance

	Informal → Formal	
	E&M	F&R
Source	49.19	50.94
Baseline	69.81	75.02
CL-SL	69.83	74.90
CL-WR	70.05	74.62
CL-ED	**70.34**	**75.41**

Analysis

Investigate which types of sentences are improved

- Procedures
  - Divide the test set based on difficulty levels
  - Compute a BLEU score of each class
  - Calculate improvements over Baseline
- Results
  - CL-SL and CL-WR degraded the quality of difficult samples
  - CL-ED improved the quality of all classes regardless of difficulty levels



(a) Sentence Length (b) Word Rarity (c) Edit Distance

[1] Platanios et al. (NAACL 19) Competence-based Curriculum Learning for Neural Machine Translation
[2] Rao and Tetreault (NAACL 18) Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer

kadotani.sora@ist.osaka-u.ac.jp