Improving Japanese-to-English Neural Machine Translation by Paraphrasing the Target Language

Tokyo Metropolitan University
Yuuki Sekizawa, Tomoyuki Kajiwara, Mamoru Komachi
sekizawa-yuuki@ed.tmu.ac.jp

Overview: vocabulary restriction in NMT

- NMT has a high computational cost
  - NMT restricts vocabulary to frequent words
  - infrequent words being treated as out-of-vocabulary (OOV)
    - substituted with a special symbol such as “<unk>”
- byte pair encoding (BPE): to split to substrings
  - do not consider word meaning
    - may change sentence meaning
- proposed method
  - **paraphrase infrequent words or phrases with frequent synonyms** from the target side of the training corpus
    - reduce OOV in output while keeping sentence meaning

Related work: reducing OOV

- Luong et al. (2015)
  - translate OOV words with a corresponding word in the source sentence using a translation dictionary
  - only use correspondence of a source word
- Sennrich et al. (2015)
  - apply BPE to source and target corpora to split OOV words into units of frequent substrings
  - split words greedily without considering their meaning

Related work: using word similarity

- Li et al. (2016)
  - substitute OOV words in training corpora with a similar in-vocabulary word as preprocessing steps
    - using cosine similarity and language model
  - lose sentence meaning and degrades the adequacy
  - might replace OOV words with similar but **non-synonymous** words since they used distributional similarity

example

- Li et al.: "internet *surfing*" → "internet snowboard"
- ours: "internet *surfing*" → "internet browser"

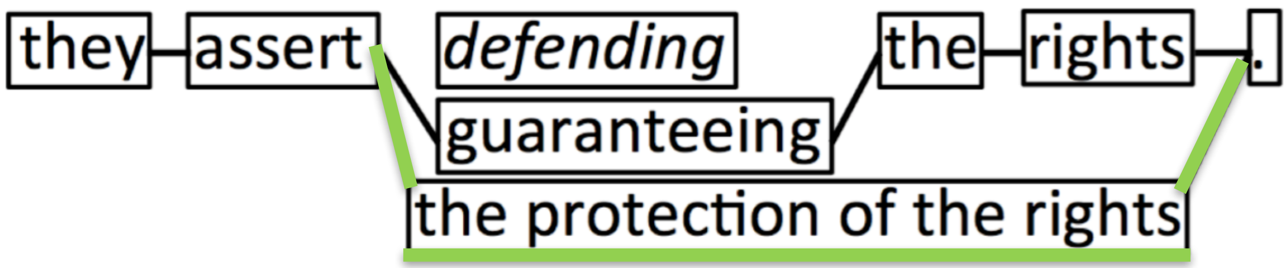
Proposed method

- paraphrase infrequent words or phrases with frequent synonyms on the target side of the training sentences

- advantage
  - consider word meaning by paraphrasing OOV
    - preserve sentence meaning while reducing OOV
  - combine related methods
  - treat NMT as a black box

Proposed method

- use a paraphrase lexicon that has paraphrase pairs annotated with a **paraphrase score (PPDBscore)**
  - combine PPDBscore and **language model score (LM score)**
    - $\lambda(\text{PPDBscore}) + (1 - \lambda)(\text{LMscore})$



OOV : *defending*

- viterbi algorithm calculates the 2-gram language model score

Proposed method: paraphrasing iteratively

original: the ***pedagogues*** had ***quarrels***.

paraphrase first round : the ***educators*** had discussions.

paraphrase second round : the teachers had discussions.

OOV : ***pedagogues***, ***quarrels***, ***educators***

multi.: ***pedagogues*** → teachers, ***quarrels*** → discussions

single: ***quarrels*** → discussions

when paraphrasing only once, ***pedagogues*** is <u>not paraphrased</u> to keep original meaning

Experiment: NMT Settings

- corpus
  - ASPEC (Japanese-to-English)
    - train : 827,503 (sentence length <= 40)
    - develop : 1,790　　test : 1,812
    - tokenizer: mecab (IPAdic) for Ja,　Moses script for En
- NMT model
  - system : OpenNMT-py (2 layers bi-LSTM)
  - batch size 64, epoch 20, embedding size 500, hidden size 500
  - dropout rate 0.3, optimizer SGD with learning rate 1.0
  - vocabulary size of source and target 30,000 respectively
- evaluation metric
  - BLEU, METEOR, a number of OOV in translated sentences

Experiment: paraphrasing settings

- paraphrase lexicon
  - PPDB 2.0 XXXL-size (Pavlick et al., 2015)
- language model
  - all sentences of the target side of ASPEC, 2-gram model

Experiment: comparison methods

- baseline: without any paraphrasing
- Luong et al.: **replace OOV by translating corresponding source word**.
- Sennrich et al.: **apply BPE** to reduce OOV
- Li et al.: **replace OOV with the most similar word** before training
- ours: **paraphrase OOV with paraphrasing lexicon**
  - a number of paraphrasing: {single, multi.}
  - a unit of paraphrasing: {word, phrase}
  - λ: {0, 0.25, 0.5, 0.75, 1}

Result: Japanese-to-English translation

method	BLEU↑	METEOR↑	OOV↓
baseline	25.70†	31.06	1,123
Luong et al. (dictionary replace)	25.87†	31.04	567
Sennrich et al. (BPE)	25.92*	31.50	**0**
Li et al. (similarity replace)	25.89*	31.10	832
proposed (multi. word, $\lambda = 0.5$)	**26.45**	**31.62**	638

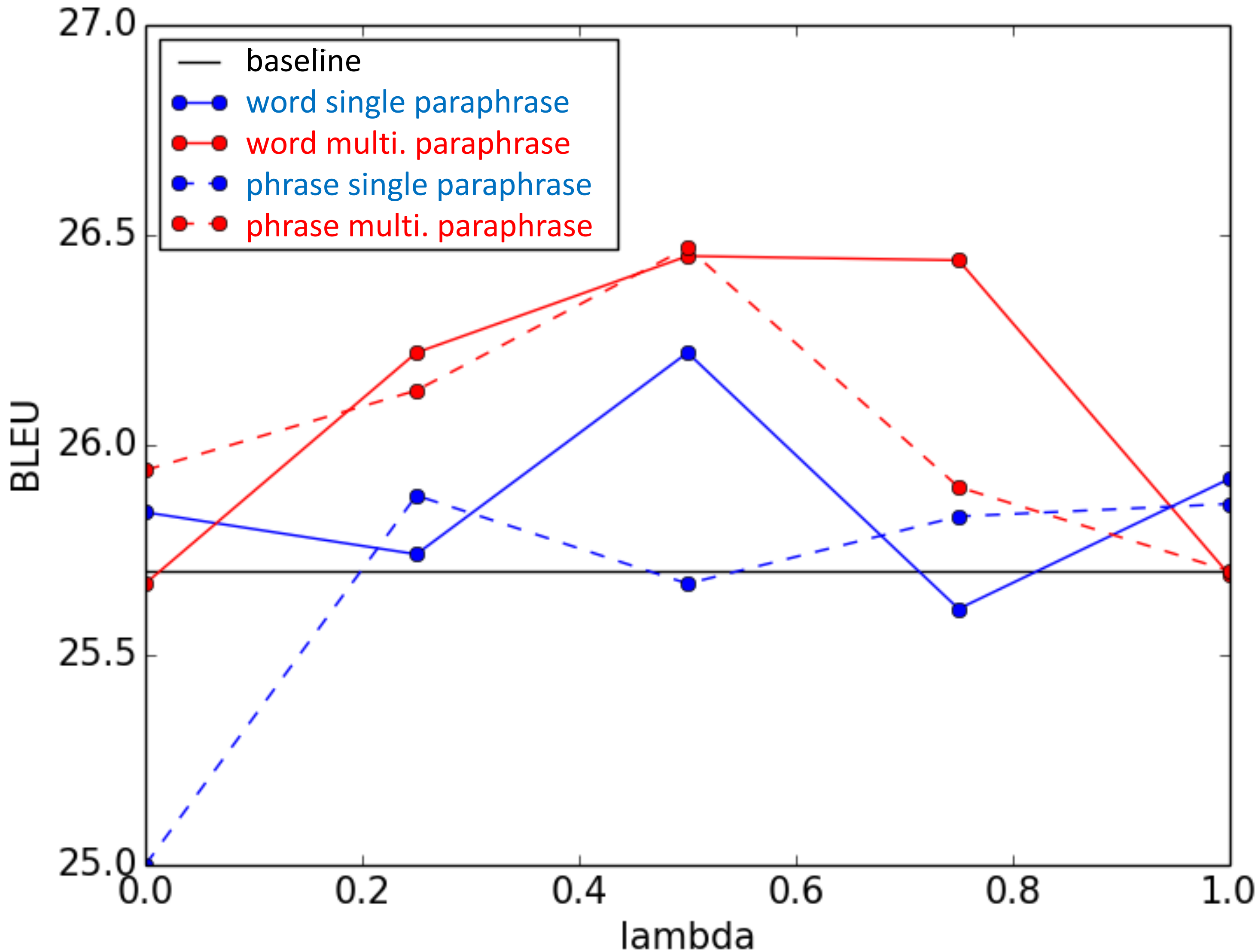
"†" and "*" indicate that the **proposed method significantly outperformed** the other methods at "$p<0.01$" and "$p<0.05$", respectively, using bootstrap resampling.

Related work reduced OOV and improved translation quality, however, our method can improve BLEU further.

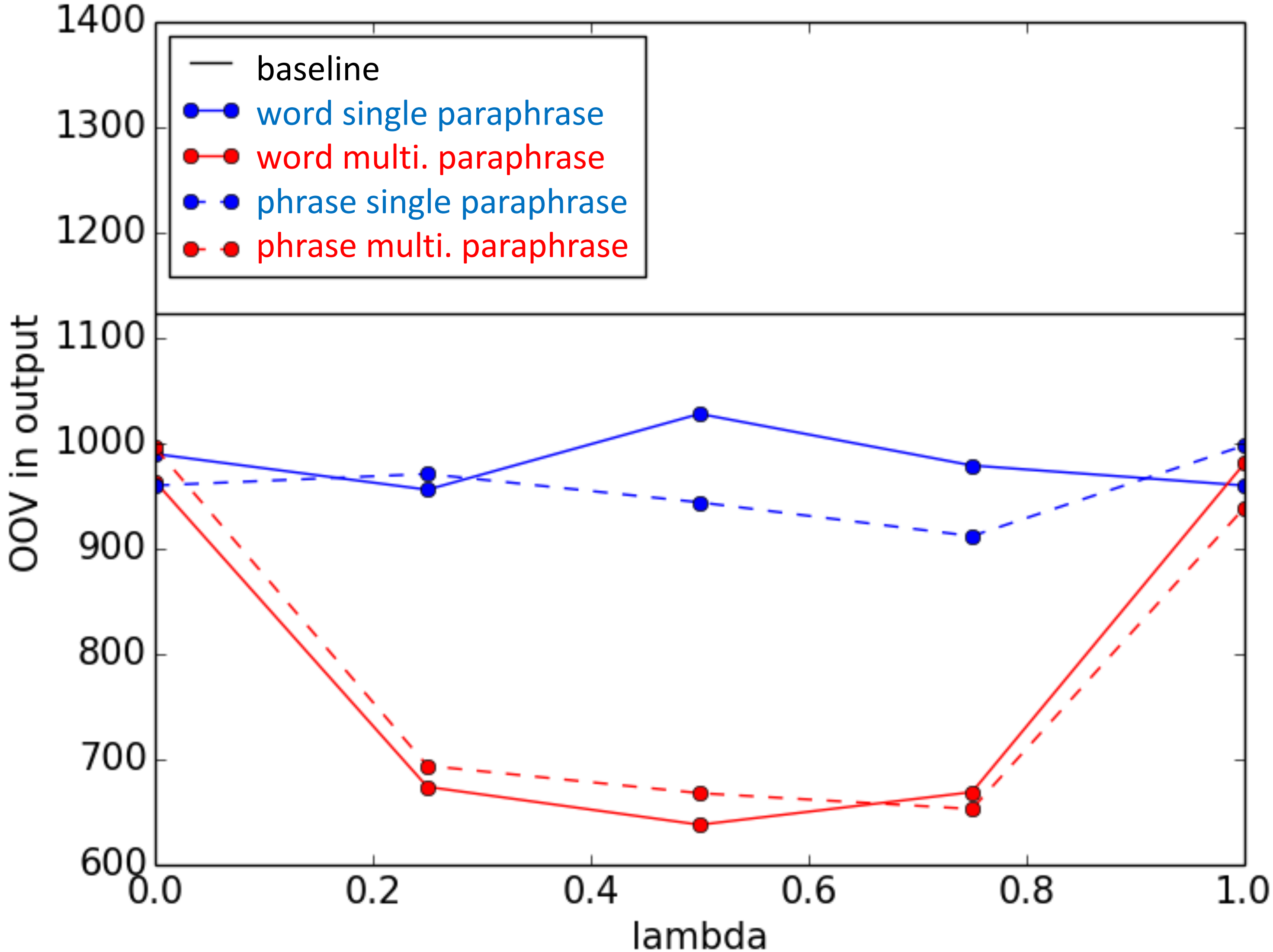
Result: translation example

method	translation
source	ロックインアンプを使用すれば, ノイズを著しく減少できることを期待できる。
reference	with the lock - in amplifier used , significant reduction of the noise is expected .
baseline	it is expected that the noise can be reduced remarkably , if the <unk> is used .
ours (multi. word, λ = 0.5)	it is expected that the noise can be remarkably decreased , if the amplifier is used .

Result: BLEU score of the proposed method



Result: a number of OOV terms in the output



Discussion

- Paraphrasing target(English) training corpus is effective.
  - Is paraphrasing **target** training corpus effective?
  - Is paraphrasing **English** training corpus effective?

Additional experiment

- We performed English to Japanese translation using PPDB: Japanese (Mizukami et al., 2014).
  - If other language pair translation performs similarly, paraphrasing **target** training corpus is effective

Using high quality lexicon is important

	En-to-Ja		Ja-to-En	
method	BLEU↑	OOV↓	BLEU↑	OOV↓
baseline	33.91	1,003	25.70†	1,123
proposed multi. word, λ = 0.5	**34.09**	**966**	**26.47**	**668**

- We can improve translation quality by paraphrasing target corpus using paraphrase lexicon.
- The degree of improvement depends on quality of lexicon.
  - English PPDB2.0 is a **supervised** regression model.
  - Japanese PPDB is an **unsupervised** model.

Proposed method comparison

Ja → En	BLEU	METEOR
non paraphrase	25.70	31.06
source paraphrase	25.94	31.02
target paraphrase	**26.45**	**31.62**
both paraphrase	25.77	31.11

En → Ja	BLEU
non paraphrase	33.91
source paraphrase	33.68
target paraphrase	**34.09**
both paraphrase	33.63

Conclusion

- **paraphrase infrequent words or phrases with frequent synonyms** from the target side of the training corpus
  - reduce OOV in output while keeping sentence meaning

- achieved a statistically significant BLEU score improvement over baselines and
  reduce the OOV rate in output sentences