Improving Neural Machine Translation by Incorporating Hierarchical Subword Features

Makoto Morishita, Jun Suzuki*, Masaaki Nagata
NTT Communication Science Laboratories

* Current affiliation is Tohoku University

Summary

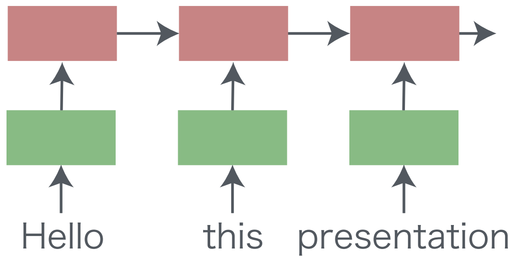
A new way to enhance embedding layer for both encoder and decoder of NMT

- use smaller subword units as additional features

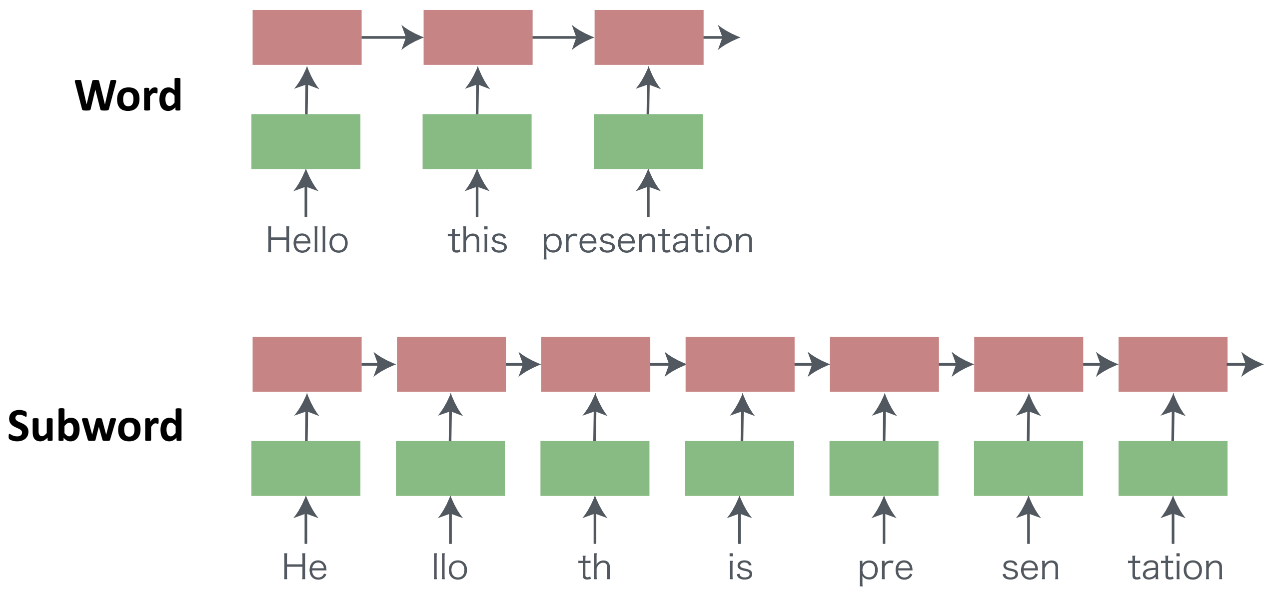
Our method can improve translation without additional computational cost

Summary

Word



Summary



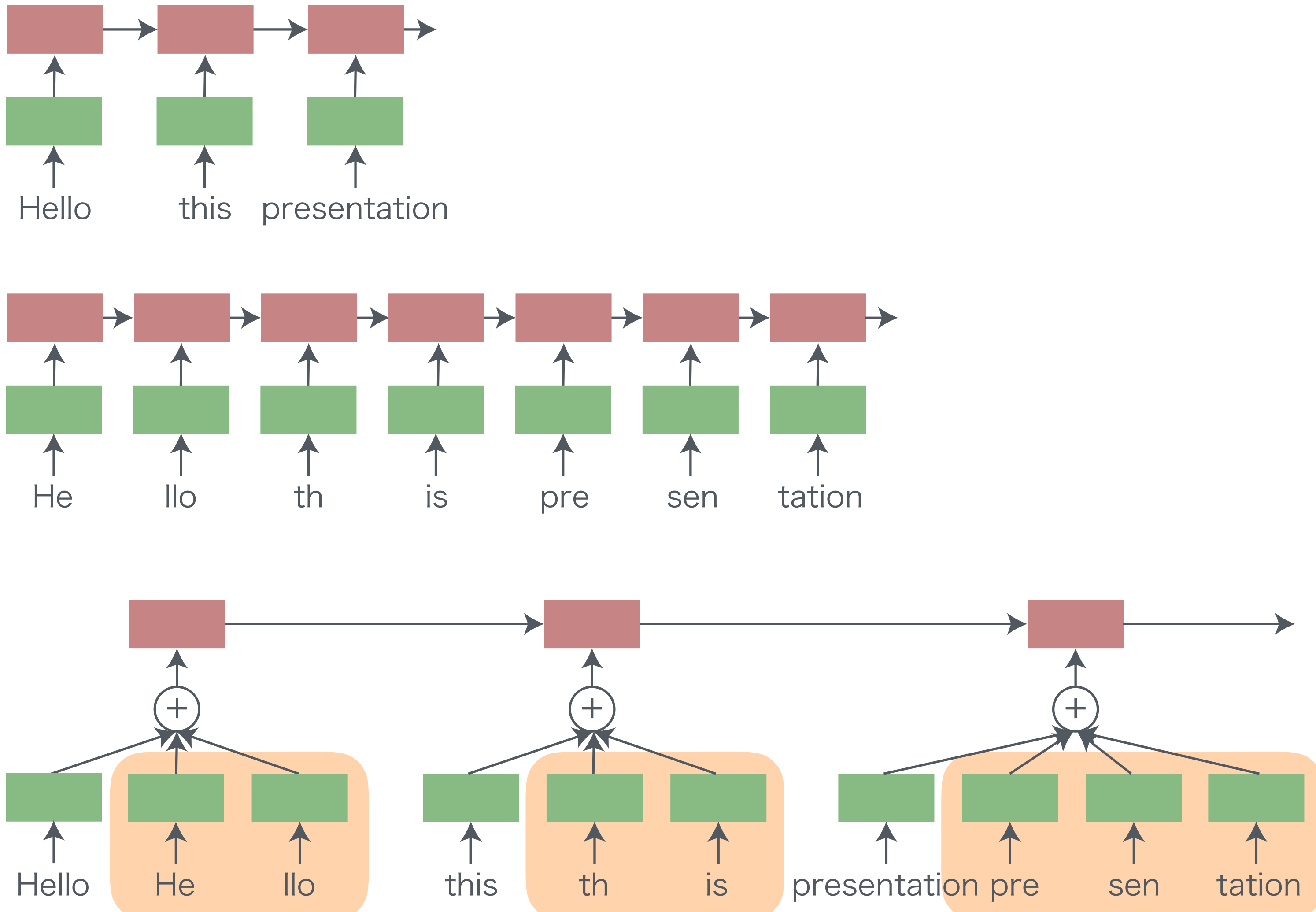
Word
Hello
this
presentation
Subword
He
llo
th
is
pre
sen
tation

Summary

Word

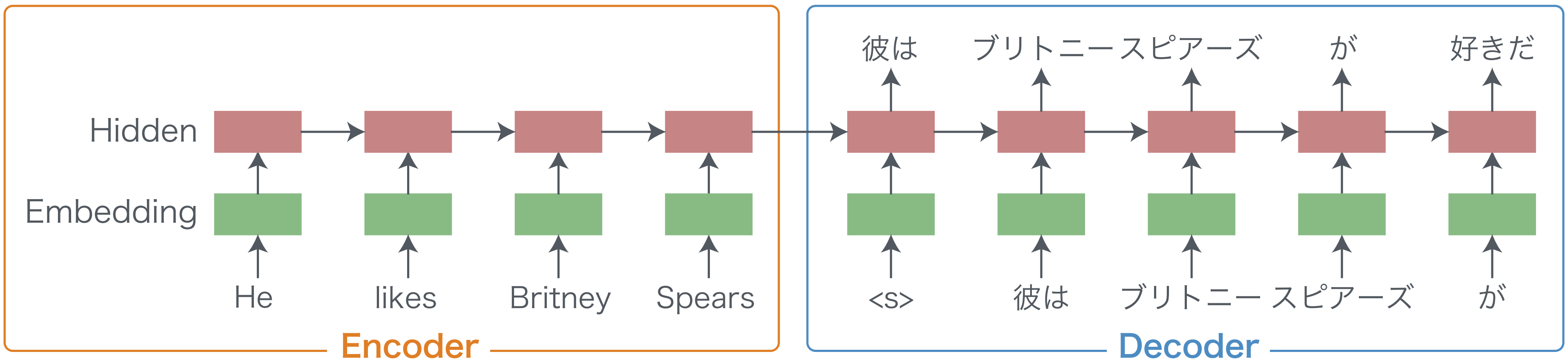
Subword

Hierarchical Subword



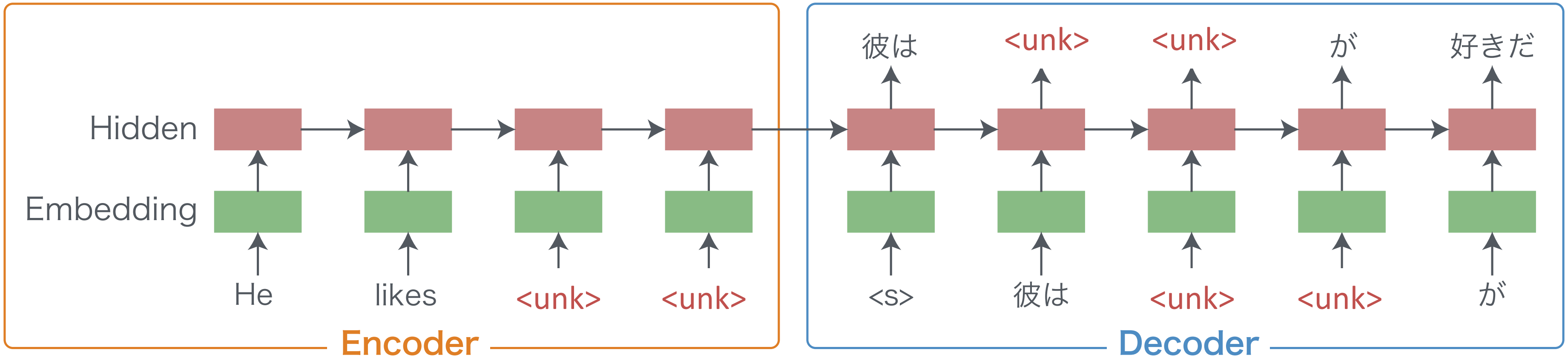
Background

Architecture of Neural Machine Translation



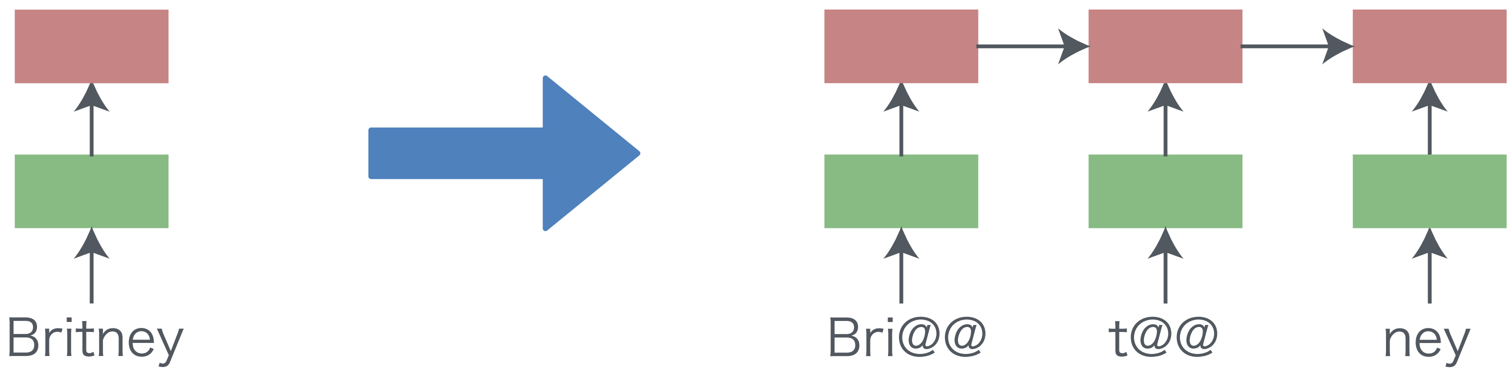
- Encoder converts a source sentence into (sequence of) vectors
- Decoder outputs a translated sentence based on the encoded vectors

Vocabulary Problem



- Traditional NMT only uses a word as a unit.
  - It cannot use the whole vocabulary.
  - We need to convert rare words into unknown word tokens.

Byte Pair Encoding (BPE)



- Split a rare word into subwords
  - Each subword is common
  - Alleviate rare words problem

"Neural Machine Translation of Rare Words with Subword Units", Sennrich et. al., ACL 2016

Pros and Cons of BPE

Pros

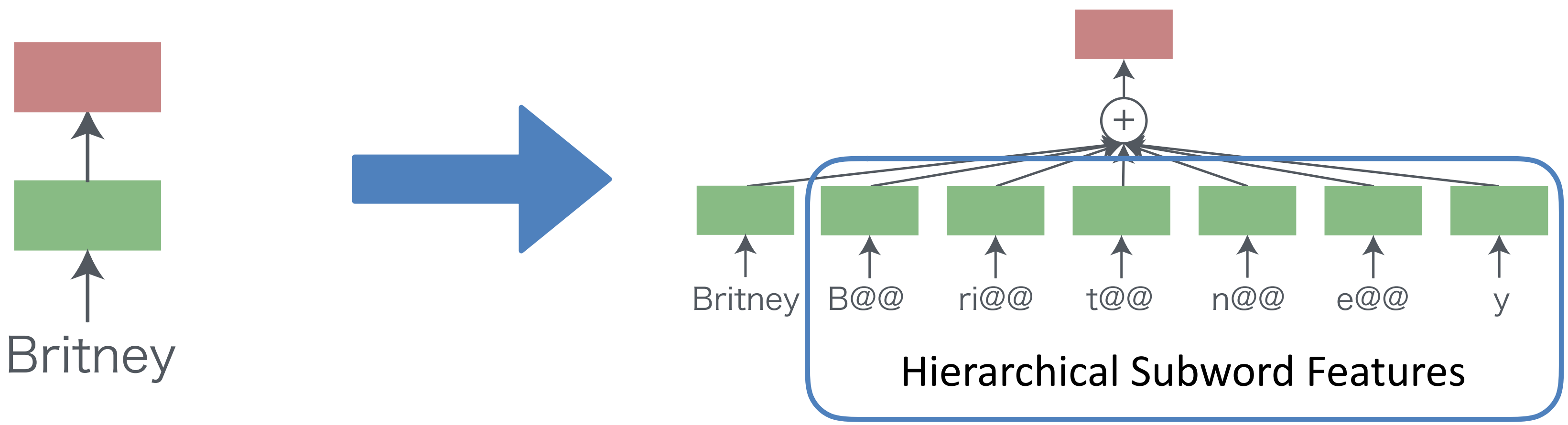
- Alleviate rare words problem
- Simple and Fast
- Fixed size of vocabulary
- Known to improve an accuracy

Cons

- Need to find appropriate unit sizes (= number of merge operations) for encoding/decoding

Proposed Method

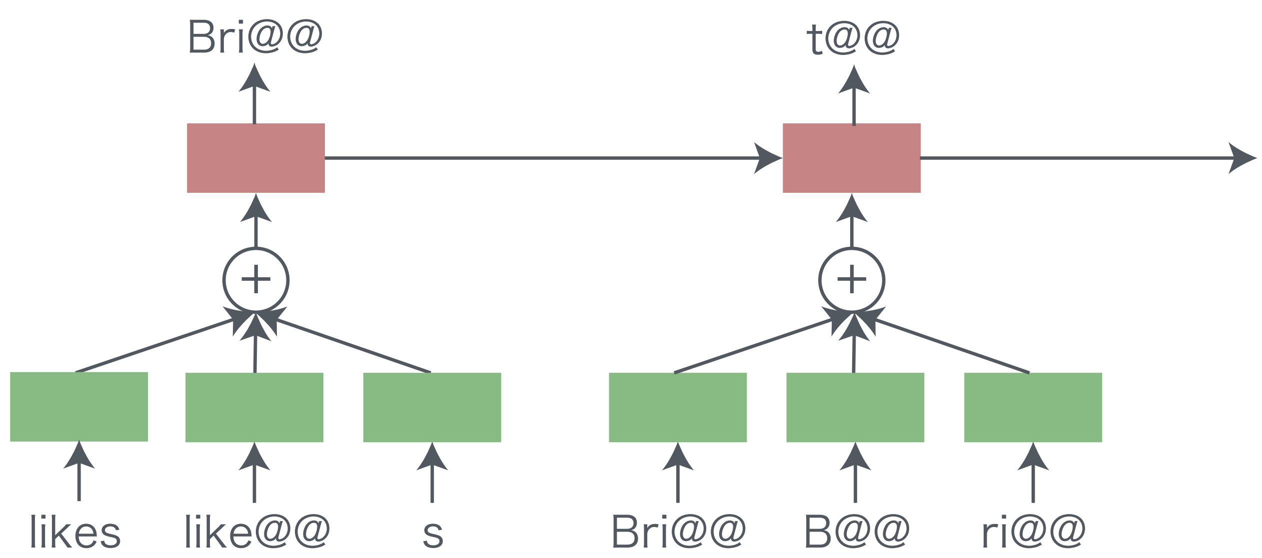
Hierarchical Subword Features



Add smaller subword units as features

- Embedding = large subword + sum of smaller subwords
- NMT can make use of several units at once

Add to Decoder Side



- It does not change an output layer.
- Hierarchical subwords can uniquely determined.

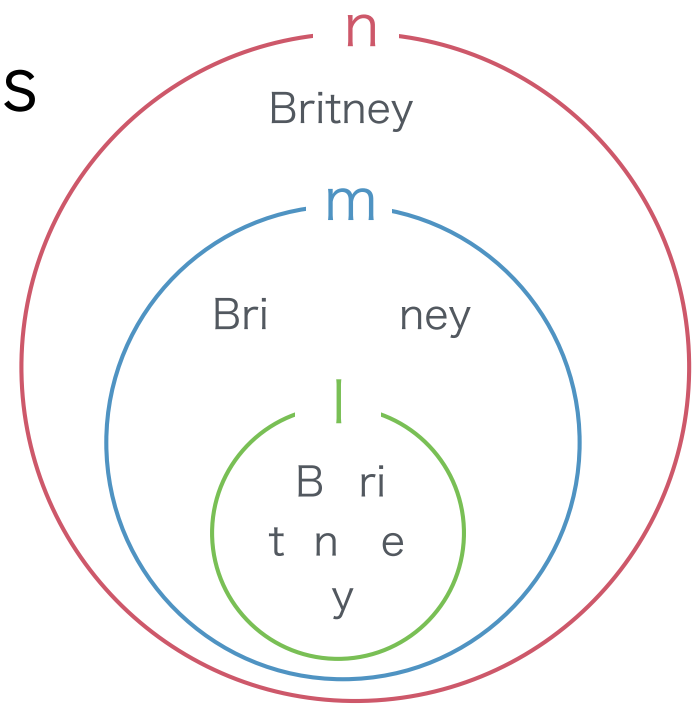
  Bri@@ → B@@ + ri@@

Add More Features

Hierarchy of BPE subwords

- Merge operations

  l < m < n

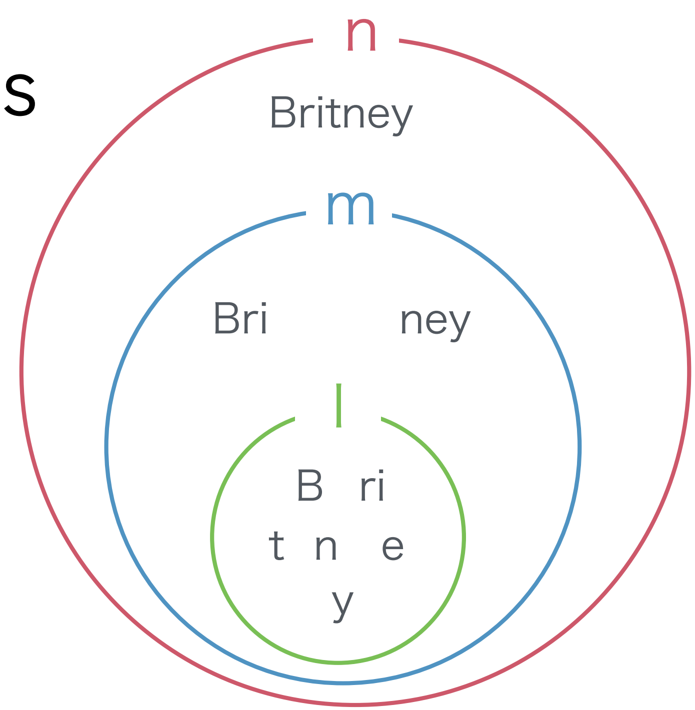


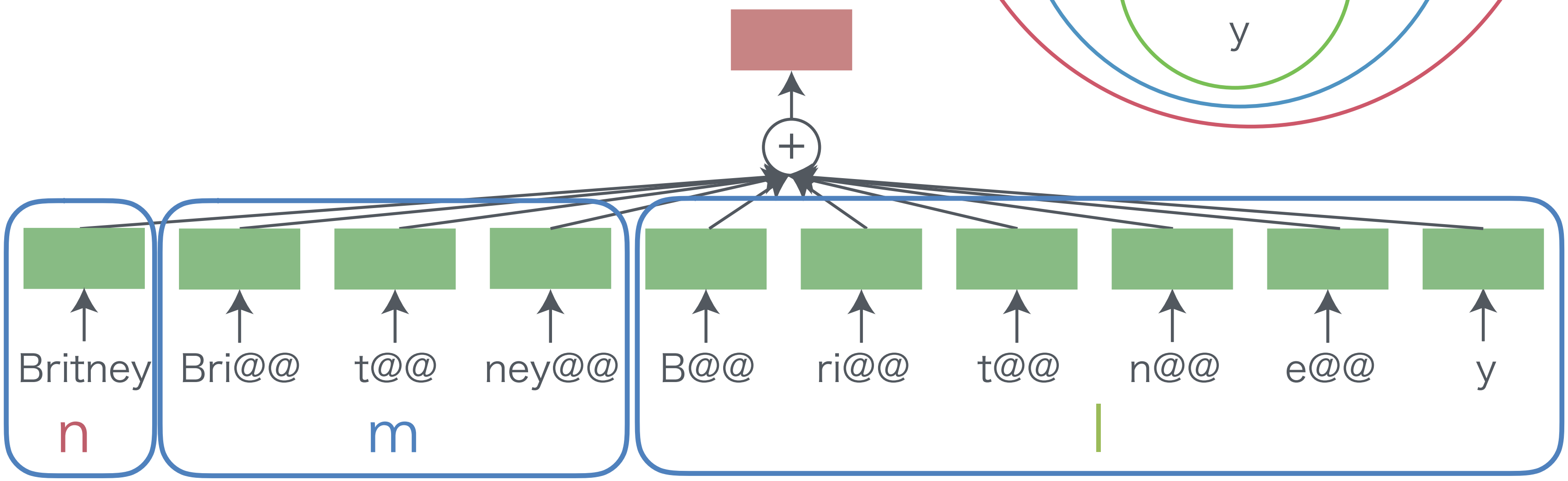
Add More Features

Hierarchy of BPE subwords

- Merge operations

  l < m < n





Implementation

One-hot (normal)

$$\boldsymbol{W}_E \begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix} =$$

Hierarchical Subword Features

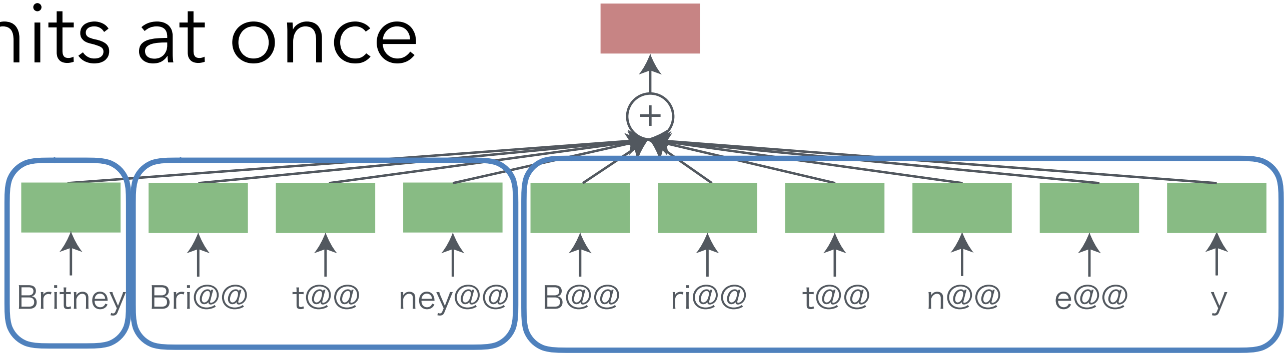
$$\boldsymbol{W}_E \begin{bmatrix}0\\1\\0\\1\\0\end{bmatrix} =$$

Multiple rows are one.

- Easy to implement!
- (Almost) No additional computational cost!

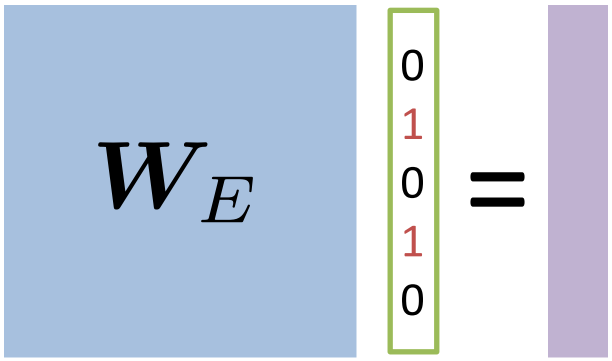
Pros of Hierarchical Subword Features

- Encoder/Decoder can use several subwords units at once



- Simple

- (Almost) No computational cost



Experiments

Research Questions

Q Does the hierarchical subword features improve the model?

Q Which part of the model should we use it?

Q Does it affect to the training speed?

Q How does it affect to the translation results?

Experimental Settings

- Corpus
  - Language: Fr-En, En-Fr
  - Training: IWSLT 2016 (TED Talk)
  - Dev: tst2014
  - Test: tst2012, tst2013

	Words	Sentences
Train	3.2M	189.3K
tst2012	30.9K	1.7K
tst2013	21.0K	1.0K
tst2014	25.0K	1.3K

Experimental Settings

- NMT model

  Encoder-decoder + attention (Luong et al., 2015)

- Vocabulary settings
  - Unit: Word level
  - Hierarchical Subword Features
    - BPE 1k and 300 vocabularies

Experimental Results

System	Averaged BLEU of 4 models	
	Fr-En	En-Fr
Baseline (BPE16k)	42.35	43.65

Experimental Results

System	Averaged BLEU of 4 models	
	Fr-En	En-Fr
Baseline (BPE16k)	42.35	43.65
Add encoder features	43.82 (+1.47)	45.32 (+1.67)

Experimental Results

System	Averaged BLEU of 4 models	
	Fr-En	En-Fr
Baseline (BPE16k)	42.35	43.65
Add encoder features	43.82 (+1.47)	45.32 (+1.67)
Add decoder features	42.55 (+0.20)	43.54 (-0.11)

Experimental Results

System	Averaged BLEU of 4 models	
	Fr-En	En-Fr
Baseline (BPE16k)	42.35	43.65
Add encoder features	43.82 (+1.47)	45.32 (+1.67)
Add decoder features	42.55 (+0.20)	43.54 (-0.11)
Add both features	43.63 (+1.28)	45.43 (+1.78)

Experimental Results

System	Averaged BLEU of 4 models	
	Fr-En	En-Fr
Baseline (BPE16k)	42.35	43.65
Add encoder features	43.82 (+1.47)	45.32 (+1.67)
Add decoder features	42.55 (+0.20)	43.54 (-0.11)
Add both features	43.63 (+1.28)	45.43 (+1.78)

Q Does the hierarchical subword features improve the model?

A Yes!

Experimental Results

System	Averaged BLEU of 4 models	
	Fr-En	En-Fr
Baseline (BPE16k)	42.35	43.65
Add encoder features	**43.82 (+1.47)**	45.32 (+1.67)
Add decoder features	42.55 (+0.20)	43.54 (-0.11)
Add both features	43.63 (+1.28)	**45.43 (+1.78)**

Q Which part of the model should we use it?

A It depends on the settings, but encoder side only or both may work well

Training Speed

System	Training time / epoch
Baseline	1050 s
Add encoder feature	1002 s
Add decoder feature	1004 s
Add both feature	1019 s

Q Does it affect to the training speed?

A No!

Example of Improved Translation

Input	J'ai répondu, "Je ne suis pas **Britney Spears**, mais tu peux peut-être me l'apprendre à moi.
Reference	I was like, "Well I'm not **Britney Spears**, but maybe you could teach me.
Baseline	I said, "I'm not **British Speney Spears**, but maybe you can teach me.
Proposed	I said, "I'm not **Britney Spears**, but maybe you can teach me.

Q How does it affect to the translation results?

A Proposed method could help to translate the rare words.

Conclusion

Hierarchical subword features improve translation accuracy!

- Simple
- (Almost) No additional computational cost
- Easy to adapt many NLP tasks.

Future work

- Try with Transformer
- Adapt to other tasks

End