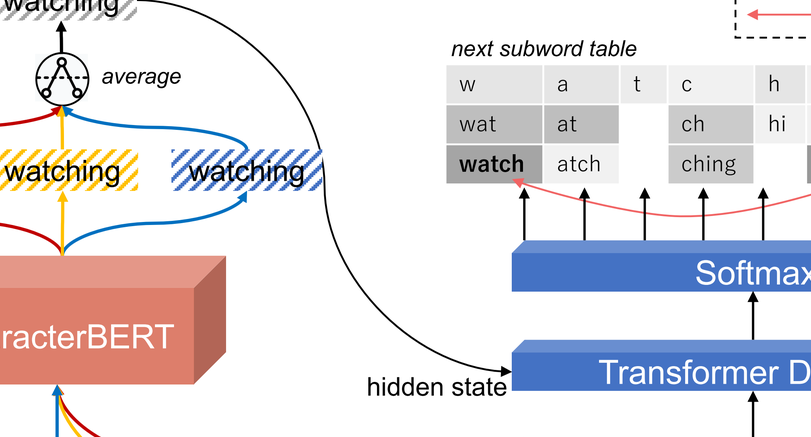
BERTSeg: BERT Based Subword Segmentation for Neural Machine Translation

Haiyue Song[†,‡], Raj Dabre[‡], Zhuoyuan Mao[†], Chenhui Chu[†], Sadao Kurohashi[†]
[†]Kyoto University [‡]NICT

Introduction

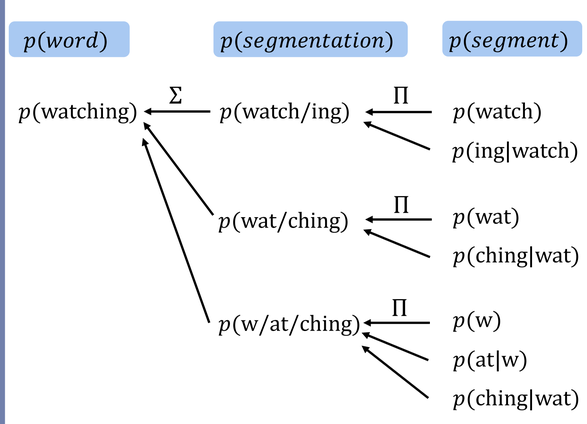
- Subword segmentation methods[1,2] rely on statistical information
- We propose:
  - **BERTSeg**, a BERT-based neural method that relys on semantic information of the target word
  - A **regularization** method for BERTSeg



Method

1 Architecture

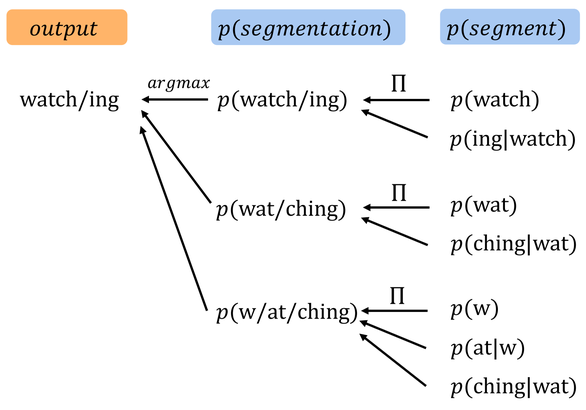
2 Train



$$\log p(\boldsymbol{x}_{1:T}|\boldsymbol{e}_{\boldsymbol{x}}) =$$

$$\log \sum_{\boldsymbol{a}_{1:\tau_a} \in \mathcal{S}(\boldsymbol{x})} \prod_{i=1}^{\tau_a} p(a_i|\boldsymbol{e}_{\boldsymbol{x}}; x_1, \ldots, x_j)$$

3 Inference



$$\boldsymbol{a} = \arg\max_{\boldsymbol{a}_{1:\tau_a} \in \mathcal{S}(\boldsymbol{x})} \prod_{i=1}^{\tau_a} p(a_i|\boldsymbol{e}_x; x_1, \ldots, x_j)$$

$$p_{sample}(\boldsymbol{a}_i) = \frac{e^{\log p(\boldsymbol{a}_i)/t}}{\sum_{\boldsymbol{a}_i \in \mathcal{S}(\boldsymbol{x})} e^{\log p(\boldsymbol{a}_i)/t}}$$

Results

1 Machine Translation

- **Higher BLEU scores** on low-resource to high-resource datasets compared with BPE[1], VOLT[2], DPE[3], BPE-dropout[4]

	ALT Asian Langs→En	IWSLT15 Vi→En	WMT16 Ro→En	WMT15 Fi→En
w/o Regularization				
BPE[1]	19.76	27.09	**32.54**	17.45
VOLT[2]	19.91	27.16	31.89	17.25
DPE[3]	19.88	27.40	29.95	16.14
BERTSeg	**20.71**	**27.80**	32.33	**17.54**
With Regularization				
BPE-dropout[4]	23.04	28.76	33.59	**18.50**
BERTSeg-Regularization	**24.68**	**30.09**	**33.82**	18.46

2 Speed

- **BERTSeg** requires about 400 seconds on large corpus during training, which is much faster than previous neural method DPE[3]

	ALT	WMT16 Ro-En
BPE[1]	4	13
VOLT[2]	960	1,747
DPE[3]	3,477	68,334
BERTSeg	58	391

3 Segmentation Examples

- High generalization ablity on ***rare*** or ***unseen*** words compared with BPE[1]

BERTSeg		BPE	
Frequent words			
official/s	officials	officials	
edit/ion	edition	edition	
use/d	used	used	
farm/er/s	far/mers	far/mers	
contribute/d	contrib/uted	contrib/uted	
normal/ly	norm/ally	norm/ally	
seven/th	sevent/h	sevent/h	
Rare words			
inter/face/s	inter/f/aces	inter/face/s	inter/f/aces
sea/side	se/as/ide	sea/side	se/as/ide
ab/normal/ly	ab/n/orm/ally	ab/normal/ly	ab/n/orm/ally
b/y/stand/er	by/st/ander	b/y/stand/er	by/st/ander
dis/comfort	disc/om/fort	dis/comfort	disc/om/fort
un/warrant/ed	un/w/arr/anted	un/warrant/ed	un/w/arr/anted
in/definitely	ind/ef/in/itely	in/definitely	ind/ef/in/itely
Unseen words			
stable/d	st/ab/led	stable/d	st/ab/led
save/r/s	sa/vers	save/r/s	sa/vers
M/illion/s	Mill/ions	M/illion/s	Mill/ions
Free/way	Fre/ew/ay	Free/way	Fre/ew/ay
M/i/s/behavior	M/is/be/hav/ior	M/i/s/behavior	M/is/be/hav/ior
m/o/u/r/n/ed	m/our/ned	m/o/u/r/n/ed	m/our/ned
M/a/d/a/m/e	Mad/ame	M/a/d/a/m/e	Mad/ame

- Subword segmentation with regularization
  - Global best N segmentations are obtained through Dynamic Programming algorithm in O(N/*og*N*T[2])

BERTSeg-Regularization Segmentation	
represent/ed	represented
represent/e/d	re/presented
re/presented	re/present/e/d

Conclusion

- **We proposed BERTSeg**, an unsupervised neural subword segmenter for NMT, together with a regularization algorithm
- **MT results** showed significant improvement over frequency-based and neural network-based methods
- **The training is efficient** even compared with non-neural methods

Future Work

- **A multilingual segmenter** using embeddings from BERT, mBERT, or character-level mBERT
- Remove the dependency on the BPE vocabulary

[1]Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. ACL.
[2]Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary Learning via Optimal Transport for Neural Machine Translation. ACL.
[3]Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation. ACL.
[4]Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and Effective Subword Regularization. ACL.