FACEBOOK

Neural Machine Translation with Byte-Level Subwords

Changhan Wang
Kyunghyun Cho
Jiatao Gu

Byte-Level Vocabularies

- Based on byte-level representation (e.g. UTF-8 encoding) of sentence strings
- Using BPE (byte-pair encoding) to segment byte stream into byte n-grams

片 | 手の | 拍手 | の | 音
片 | 手 E3 81 | AE | 拍 | 手 E3 81 | AE | E9 9F | B3

Compacter than pure characters

- Pure bytes: maximum 256 possible bytes
- Byte-level BPE (BBPE): any size >= 257, can be compacter than pure characters!
- BBPE has fewer rare symbols and shorter tokenized sentences (runs faster)

Generic and having no OOV tokens

Any sentence strings can be represented by bytes. BBPE contains all bytes and has no OOV tokens.

Example: BPE vs. BBPE

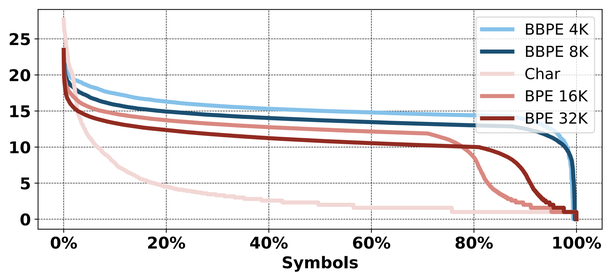
Original		質問して__証明と証拠を求めましょう	Ask__questions,__demand__proof,__demand__evidence.
Byte		E8 B3 AA E5 95 8F E3 81 97 E3 81 A6 E2 96 81 E8 A8 BC E6 98 8E E3 81 A8 E8 A8 BC E6 8B A0 E3 82 92 E6 B1 82 E3 82 81 E3 81 BE E3 81 97 E3 82 87 E3 81 86	41 73 6B E2 96 81 71 75 65 73 74 69 6F 6E 73 2C E2 96 81 64 65 6D 61 6E 64 E2 96 81 70 72 6F 6F 66 2C E2 96 81 64 65 6D 61 6E 64 E2 96 81 65 76 69 64 65 6E 63 65 2E
BBPE	1K	E8 B3 AA E595 8F しE381 A6 __E8 A8 BC 明 E381 A8 E8 A8 BC E6 8B A0 をE6 B1 82 めE381 BE しょう	A s k __quest ions , __dem and __pro of , __dem and __ev idence .
	2K	E8 B3 AA 問 しE381 A6 __E8 A8BC 明 E381 A8 E8 A8BC E68B A0 をE6 B1 82 めE381 BE しょう	A s k __qu est ion s , __d em and __pro o f , __d em and __e v id ence .
	4K	E8 B3 AA 問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82 めE381 BE しょう	As k __quest ions , __d em and __pro of , __d em and __ev id ence .
	8K	E8 B3 AA問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82めE381 BE しょう	As k __questions , __demand __pro of , __demand __evidence .
	16K	E8 B3 AA問 しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82めE381 BE しょう	As k __questions , __demand __proof , __demand __evidence .
	32K	E8 B3 AA問しE381 A6 __E8 A8BC 明E381 A8 E8 A8BC 拠 をE6 B1 82 めE381 BE しょう	As k __questions , __demand __proof , __demand __evidence .
CHAR		質 問 し て __ 証 明 と 証 拠 を 求 め ま し ょ う	A s k __ q u e s t i o n s , __ d e m a n d __ p r o o f , __ d e m a n d __ e v i d e n c e .
BPE	16K	質問 して __ 証明 と 証拠 を求 め ましょう	As k __questions , __demand __pro of , __demand __evidence .
	32K	質問 して __証明 と 証拠 を求め ましょう	As k __questions , __demand __proof , __demand __evidence .

Embedding Contextualization

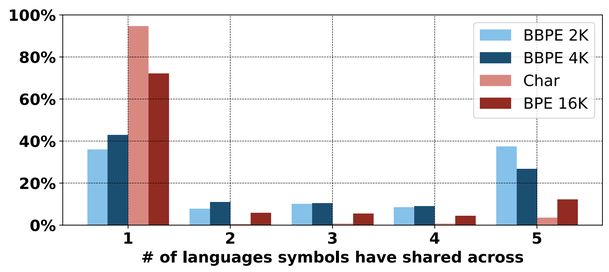
BBPE symbols are finer-grained and more generic. Contextualized embeddings (via convolution/RNN) help better disambiguation.

Qualitative Comparison: BPE vs. BBPE

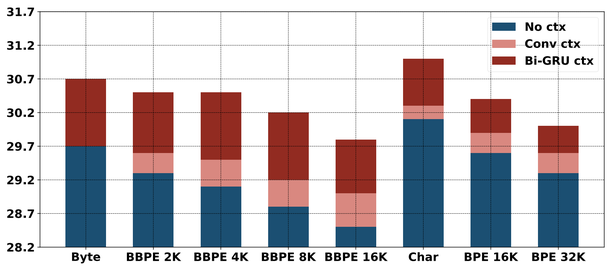
BBPE is less long tail distributed



BBPE has more cross-lingual sharing



Importance of Embedding Contextualization



Experimental Results

Noisy Character Set: En-De

		Test BLEU	Params
T_{base}	Byte+	26.59	45M
	BBPE 2K+	26.98	47M
	BBPE 4K+	27.08	47M
	Char+	26.73	47M
	BPE 32K	27.31	61M
	BPE 32K+	**27.41**	62M
	BPE 37K*	27.3	65M
T_{big}	Byte+	26.94	181M
	BBPE 2K+	**28.78**	183M
	BBPE 4K+	28.27	185M
	Char+	27.24	185M
	BPE 32K	28.36	210M
	BPE 32K+	28.77	215M
	BPE 37K*	28.4	213M

3.4K character set with a lot of non-En / non-De ones.

Character-Rich Language: Ja-En

		KFTT	TED	JESC	All
# of train samples		440K	223K	2.8M	3.5M
# of test samples		1.2K	8.5K	2K	11.7K
Michel et.al. (2018)		20.77	13.25	18.00	-
T_{base}	Byte+	23.12	15.14	15.69	16.27
	BBPE 4K+	**24.15**	15.59	16.10	16.80
	Char+	23.67	15.26	15.68	16.43
	BPE 16K+	23.63	**16.15**	**16.18**	**17.19**
T_{big}	Byte+	23.68	16.08	16.29	17.46
	BBPE 4K+	23.88	**19.0**	**17.93**	**19.58**
	Char+	23.71	16.69	17.01	18.33
	BPE 16K+	**24.08**	18.34	17.89	19.14

7.9K character set. 99.99% tokens covered by the top 2.4K characters.

Multilingual Setting: Many-to-En

		Ar	De	He	It	Az	Be	Gl	Sk	All	Params
# of train examples		213K	167K	211K	203K	5.9K	4.5K	10K	61K	5.1M	
# of test examples		6K	4.5K	5.5K	5.6K	0.9K	0.7K	1K	2.4K	165K	
Aharoni et al. 19		25.93	28.87	30.19	32.42						
Neubig & Hu 18						11.7	18.3	29.1	28.3		
T_{base}	Byte+	31.13	35.98	36.77	38.36	14.64	**25.12**	35.12	33.08	30.38	45M
	Char+	**31.52**	**36.73**	**36.85**	**38.62**	**15.40**	24.90	**35.44**	**33.31**	**30.75**	51M
T_{base}	BBPE 2K+	**30.79**	**35.53**	**36.27**	**37.82**	13.64	24.70	**34.17**	**32.83**	**29.91**	46M
	BBPE 4K+	30.64	34.93	36.07	37.62	**13.76**	**24.84**	33.90	32.12	29.74	47M
	BPE 16K	29.70	34.35	34.47	37.02	13.28	24.61	33.55	31.72	29.00	53M
	BPE 16K+	30.20	34.97	35.55	37.49	12.65	23.66	33.95	32.16	29.62	54M
	BPE 32K	29.02	34.08	34.18	36.63	12.56	22.48	32.33	31.26	28.81	61M
	BPE 32K+	29.87	34.64	35.26	37.43	12.35	22.05	33.62	31.61	29.43	62M

58 source languages to En. 10.8K character set.

Transfer Learning on Unseen Characters

		Train	Finetune	BLEU
T_{flores}	BPE 5K*	Si-En	-	7.2
	BBPE 4K+	Si-En	-	7.1
T_{flores}	BBPE 4K+	X-En	-	0.3
	BBPE 4K+	X-En	enc	8.3
	BBPE 4K+	X-En	enc, dec	8.1
	BBPE 4K+	X-En	embed, enc	**9.0**
	BBPE 4K+	X-En	all	8.6

Pre-training on multilingual many-to-En, and finetuning on Si-En. Si characters are unseen in the pre-trained model.

References

Costa-jussà, Marta R., Carlos Escolano, and José AR Fonollosa (2017). "Byte-based neural machine translation". *In: Proceedings of the First Workshop on Subword and Character Level Models in NLP*

Chung, Junyoung, Kyunghyun Cho, and Yoshua Bengio (2016). "A Character-level Decoder without Explicit Segmentation for Neural Machine Translation". *In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1693-1703.*

Lee, Jason, Kyunghyun Cho, and Thomas Hofmann (2017). "Fully character-level neural machine translation without explicit segmentation". *Transactions of the Association for Computational Linguistics 5 (2017): 365-378.*

Take a photo to learn more: