A Call for Prudent Choice of Subword Merge Operations in NMT

Shuoyang Ding
Adithya Renduchintala
Kevin Duh

Subword methods are a category of preprocessing step that segment rare words into more frequently observed "subword units". For example, the German word "Gesundheitsforschungsinstitute" might be segmented into "Gesundheits forsch ungsinstitu ten".
We analyze the interaction between the number of subword merge operations and various aspects.

Dataset

- **Low Resource (main results)**: IWSLT 2016 dataset, which covers translation of English from and into Arabic (ar), Czech (cs), French (fr) and German (de).
- **Language Analysis**: A subset of TED Corpus (Qi et al. 2018), covering Brazilian Portuguese (pt), Hebrew (he), Russian (ru), Turkish (tr), Polish (pl) and Hungarian (hu), paired with English.
- **High resource**: WMT 2017 Russian-English shared task, with WMT 2012-2016 testsets as devset.

Below is a summary of findings in each aspect. All analysis are done on **low-resource dataset** unless otherwise specified.

Analysis on Architecture

- **For Transformer-based architectures**, we recommend the sweep be concentrated in the 0–4k range. The performance difference between the optimal (normally 0–1k) and the worst configuration (normally 16k/32k) is generally 3–4 BLEU points.
- **For LSTM-based architectures**, we find no typically optimal BPE merge operation setup and therefore urge future work to sweep over 0–32k to the extent possible. On the other hand, the performance variation between BPE size is milder than that of Transformer architecture.

Analysis on Joint/Separate BPE

This is the comparison where we either build a **joint** BPE model for both sides of the language pair, or build models **separately** for each side.
The analysis revealed no significant performance differences between joint BPE and separate BPE. Therefore, we recommend BPE sweep be conducted with either of these settings.

Analysis on Languages

This is a regression analysis done on both **low-resource** and **language analysis dataset**. We find the performance variance with regard to BPE size significantly increase when translating into fusional languages (such as English or French) or when translating from agglutinative languages (such as Turkish)

Analysis on High Resource Setting

This is done on the **high resource dataset**. We find that 16k/32k is still optimal under this setup.

Analysis on Random Seed

We ran the **ar-en** and **en-ar** experiments multiple times with **shallow-transformer** architecture and find our conclusions still hold with different random seeds.

JOHNS HOPKINS UNIVERSITY

Subword methods are **not** something one should plug-and-forget when building NMT systems. **Configure them carefully**, or risk losing **3–4 BLEU points** of performance in some settings!

Take a picture to download the full paper

Architecture Setup

	bi-dir	d_{enc}	d_{dec}	d_{emb}	l	N_h	N_p
shallow-transformer	N/A	512	512	512	2	4	18.8M
deep-transformer	N/A	512	512	512	6	4	39.8M
tiny-lstm	no	256	256	256	1	1	5.6M
shallow-lstm	yes	384	384	384	2	1	16.4M
deep-lstm	yes	384	384	384	6	1	35.3M

Transformer Results

		0	0.5k	1k	2k	4k	8k	16k	32k	δ
deep-transformer	**ar-en**	30.3	30.8	30.6	30.5	30.4	29.8	28	27.5	3.3
	cs-en	24.6	23.3	23.0	22.7	21.2	22.6	20.6	21.0	4.0
	de-en	28.1	28.6	28.0	28.4	27.7	27.5	26.7	25.2	3.4
	fr-en	28.8	29.8	29.6	29.3	28.7	28.5	27.5	26.6	3.2
	en-ar	12.6	13.0	12.1	12.3	11.8	11.3	10.7	10.6	2.4
	en-cs	17.3	17.1	16.7	16.4	16.1	15.6	14.7	13.8	3.5
	en-de	26.1	27.4	27.4	26.1	26.3	26.1	25.8	23.9	3.5
	en-fr	25.2	25.6	25.3	25.5	25.3	24.7	24.1	22.8	2.8
shallow-transformer	**ar-en**	26.4	27.9	28.7	28.5	28.6	27.7	26.2	25.5	3.2
	cs-en	22.4	22.6	22.3	21.8	21.7	21.1	21.1	20.1	2.5
	de-en	25.5	27.4	27.1	27.3	27.1	25.9	24.6	23.7	3.7
	fr-en	26.3	28.0	28.9	28.0	28.0	27.4	26.1	26.1	2.7
	en-ar	11.7	11.2	11.5	11.0	11.3	10.5	9.5	9.0	2.7
	en-cs	16.4	16.7	16.0	16.2	14.4	14.2	13.9	13.9	2.8
	en-de	23.8	25.7	25.4	25.3	25.2	24.3	24.1	22.1	3.6
	en-fr	23.5	24.7	25.1	24.6	24.5	23.8	22.7	22.1	3.0

LSTM Results

		0	0.5k	1k	2k	4k	8k	16k	32k	δ
tiny-lstm	**ar-en**	20.6	22.1	22.4	23.0	24.1	24.2	24.2	24.0	3.6
	cs-en	17.8	19.1	18.8	19.0	19.2	19.5	20.7	19.1	2.9
	de-en	21.1	22.5	23.2	23.1	23.1	23.1	23.6	23.0	2.5
	fr-en	21.8	25.3	25.3	25.4	25.1	25.3	25.1	24.7	3.6
	en-ar	8.5	8.7	9.3	8.8	8.8	8.6	8.8	8.8	0.8
	en-cs	11.5	12.3	13.7	13.2	13.0	14.1	14.4	13.2	2.9
	en-de	18.2	20.8	21.4	21.1	21.9	21.6	21.0	21.6	3.7
	en-fr	19.9	20.4	20.7	21.8	21.3	21.0	21.3	21.3	1.7
shallow-lstm	**ar-en**	27.5	27.2	27.1	27.6	27.4	26.7	27.5	26.3	1.3
	cs-en	22.2	22.2	22.2	22.9	22.7	23.0	22.8	21.6	1.4
	de-en	25.7	25.9	26.0	25.9	26.4	26.3	26.1	26.5	0.8
	fr-en	27.6	26.7	27.7	28.4	27.9	27.7	28.5	27.5	1.8
	en-ar	11.0	11.0	10.7	10.4	10.6	10.6	10.4	10.1	0.9
	en-cs	16.1	15.7	15.8	15.3	15.8	15.5	15.8	15.6	0.8
	en-de	24.9	25.1	23.9	24.2	25.4	25.2	25.5	25.0	1.6
	en-fr	24.3	23.8	23.7	24.2	23.5	24.1	23.9	23.0	1.3
deep-lstm	**ar-en**	21.2	25.7	27.2	27.1	25.6	24.8	25.1	22.9	4.3
	cs-en	19.8	22.0	18.5	21.1	20.9	21.2	20.3	15.8	6.2
	de-en	25.7	25.2	24.9	24.1	24.5	23.5	23.5	23.1	2.6
	fr-en	25.6	26.8	27.1	26.0	26.9	25.6	17.9	22.8	9.2
	en-ar	10.9	10.2	10.3	7.5	9.5	9.4	7.2	8.0	3.7
	en-cs	13.7	14.6	15.3	14.6	12.2	12.6	11.9	12.6	3.4
	en-de	22.4	24.9	23.6	23.9	22.4	24.0	24.3	23.4	2.5
	en-fr	23.1	22.9	23.5	23.1	22.2	22.0	18.0	20.0	5.5

High Resource Results

	0	0.5k	1k	2k	4k	8k	16k	32k	δ
ru-en	29.3	30.4	30.0	30.3	30.6	30.9	31.0	30.9	1.7
en-ru	28.0	29.1	29.1	29.5	29.5	29.8	30.0	30.0	2.0

The Center For Language and Speech Processing
at the Johns Hopkins University