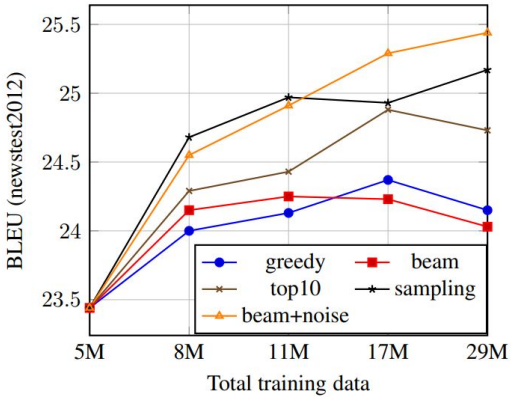
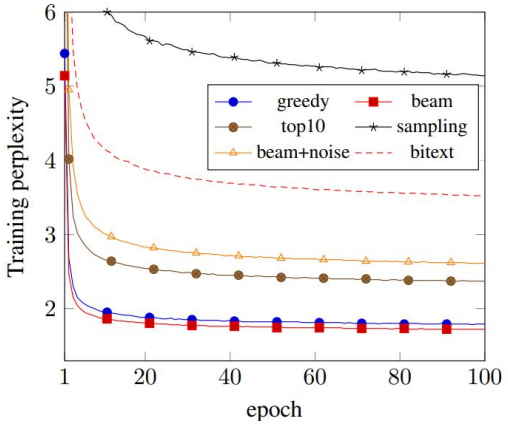
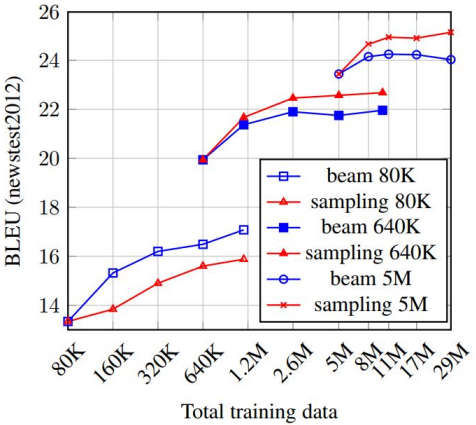
Understanding Back-Translation at Scale



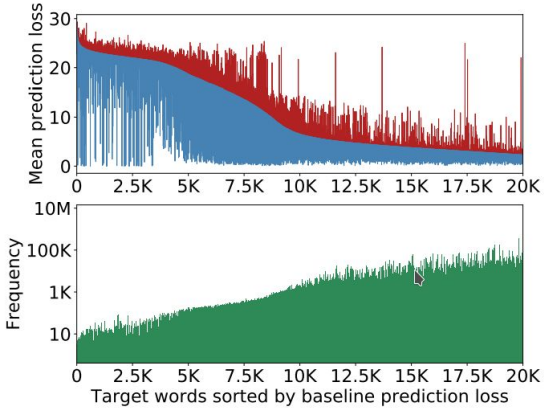


source	Diese gegenstzlichen Auffassungen von Fairness liegen nicht nur der politischen Debatte zugrunde.
reference	These competing principles of fairness underlie not only the political debate.
beam	These conflicting interpretations of fairness are not solely based on the political debate.
sample	*Mr President,* these contradictory interpretations of fairness are not based solely on the political debate.
top10	Those conflicting interpretations of fairness are not solely at the heart of the political debate.
beam+noise	conflicting BLANK interpretations BLANK are of not BLANK based on the political debate.

	En–De	En–Fr
a. Gehring et al. (2017)	25.2	40.5
b. Vaswani et al. (2017)	28.4	41.0
c. Ahmed et al. (2017)	28.9	41.4
d. Shaw et al. (2018)	29.2	41.5
DeepL	33.3	**45.9**
Our result	**35.0**	45.6
detok. sacreBLEU[3]	*33.8*	*43.8*



Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation



System	Difficulty criterion	Context	Similarity	De-En test2014	De-En test2015	De-En test2016	De-En test2017	En-De test2014	En-De test2015	En-De test2016	En-De test2017
BASELINE †				26.7	27.6	32.5	28.1	21.2	23.3	28.0	22.4
RANDOM †				28.7	29.7	36.3	30.8	24.0	26.0	30.7	24.8
FREQ	TOKENS	EMB		30.0	30.8	37.6	31.7	24.4	26.3	31.5	25.6
PREDLOSS	SWORDS	MATCH		29.1	30.1	36.9	31.0	23.8	26.2	28.8	23.2
PREDLOSS	TOKENS	MATCH		29.7	30.6	37.6	31.8	24.3	27.4	31.6	25.5
PREDLOSS	TOKENS	EMB		29.9	30.8	37.7	31.9	24.5	**27.5**	31.7	25.6
PREDLOSS	SENTENCE	EMB		24.9	25.5	30.1	26.2	22.0	24.6	27.9	22.5
MEANPREDLOSS	TOKENS	EMB		**30.2**	**31.4**	**37.9**	**32.2**	**24.4**	27.2	**31.8**	**25.6**

	Size	2014	2015	2016	2017
Baseline	4.5M	26.7	27.6	32.5	28.1
+ synthetic (1:1)	9M	28.7	29.7	36.3	30.8
+ synthetic (1:4)	23M	29.1	30.0	36.9	31.1
+ synthetic (1:10)	50M	22.8	23.6	29.2	23.9

	Size	2014	2015	2016	2017
Baseline	4.5M	21.2	23.3	28.0	22.4
+ synthetic tgt	9M	22.4	25.3	29.8	23.7
+ synthetic src	9M	24.0	26.0	30.7	24.8

	Size	2014	2015	2016	2017
Baseline	2.25M	24.3	24.9	29.5	25.6
+ synthetic	4.5M	26.0	26.9	32.2	27.5
+ ground truth	4.5M	26.7	27.6	32.5	28.1

Improving Neural Machine Translation Models with Monolingual Data

a lot of monodata

dataset	sentences
$WMT_{parallel}$	4 200 000
$WIT_{parallel}$	200 000
WMT_{mono_de}	160 000 000
WMT_{synth_de}	3 600 000
WMT_{mono_en}	118 000 000
WMT_{synth_en}	4 200 000

dataset	sentences
WIT	160 000
SETimes	160 000
$Gigaword_{mono}$	177 000 000
$Gigaword_{synth}$	3 200 000

quality comparison when +mono, +synthetic

name	training instances	BLEU newstest2014 single	BLEU newstest2014 ens-4	BLEU newstest2015 single	BLEU newstest2015 ens-4
syntax-based (Sennrich and Haddow, 2015)		22.6	-	24.4	-
Neural MT (Jean et al., 2015b)		-	-	22.4	-
parallel	37m (parallel)	19.9	20.4	22.8	23.6
+monolingual	49m (parallel) / 49m (monolingual)	20.4	21.4	23.2	24.6
+synthetic	44m (parallel) / 36m (synthetic)	**22.7**	**23.8**	**25.7**	**26.5**

	name	fine-tuning data	fine-tuning instances	BLEU tst2013	BLEU tst2014	BLEU tst2015
	NMT (Luong and Manning, 2015) (single model)			29.4	-	-
	NMT (Luong and Manning, 2015) (ensemble of 8)			31.4	27.6	30.1
1	parallel	-	-	25.2	22.6	24.0
2	+synthetic	-	-	26.5	23.5	25.5
3	2+WIT_{mono_de}	$WMT_{parallel}$ / WIT_{mono}	200k/200k	26.6	23.6	25.4
4	2+WIT_{synth_de}	WIT_{synth}	200k	28.2	24.4	26.7
5	2+$WIT_{parallel}$	WIT	200k	**30.4**	**25.9**	**28.4**

6 BLEU difference in back-translation quality leads to a 0.6–0.7 BLEU difference in translation quality

back-translation	BLEU DE→EN 2015	BLEU EN→DE 2014	BLEU EN→DE 2015
none	-	20.4	23.6
parallel (greedy)	22.3	23.2	26.0
parallel (beam 12)	25.0	23.8	26.5
synthetic (beam 12)	28.3	23.9	26.6
ensemble of 3	-	24.2	27.0
ensemble of 12	-	24.7	27.6

system	BLEU WMT	BLEU IWSLT
parallel	20.1	21.5
+synthetic	20.8	21.6
PBSMT gain	+0.7	+0.1
NMT gain	+2.9	+1.2