test set	BLEU
Random 1000 samples	11.85
With idioms	6.35
Blacklist triggered	5.64
Blacklist not triggered	6.44