Table 2: Effect of the number of full-view vector size N on One Billion Word benchmark language model. The preview window width is fixed to 256 in this table. We omitted the ratio of approximated $\tilde{Z}$ and real Z, because the ratio is over 0.997 for all cases in the table. The multiplication ratio is to full-softmax, including the overhead of $\mathbf{V}^T \times \mathbf{h}$.

V	N	NLL (full/SVD)		Top-1	Top-5	Top-100	Top-500	Mult. ratio
8K	1024	2.685	2.698	9.98	49.81	99.36	469.48	0.493
	2048	2.685	2.687	9.99	49.99	99.89	496.05	0.605
80K	4096	3.589	3.6051	10.00	49.94	99.85	497.73	0.195
	8192	3.589	3.591	10.00	49.99	99.97	499.56	0.240
400K	16384	3.493	3.495	10.00	50.00	100.00	499.90	0.171
	32768	3.493	3.495	10.00	50.00	100.00	499.98	0.201
800K	32768	4.688	4.718	10.00	49.99	99.96	499.99	0.168
	65536	4.688	4.690	10.00	49.99	99.96	499.89	0.200

Table 3: SVD-softmax on machine translation task. The baseline perplexity and BLEU score are 10.57 and 21.98, respectively.

W	N	Perplexity	BLEU
200	5000	10.57	21.99
	2500	10.57	21.99
	1000	10.58	22.00
100	5000	10.58	22.00
	2500	10.59	22.00
	1000	10.65	22.01
50	5000	10.60	22.00
	2500	10.68	21.99
	1000	11.04	22.00

The preview window width and the number of full-view vectors were selected in the powers of 2. The results were computed on randomly selected 2,000 consecutive frames.

Table 2 shows the experimental results. With a fixed hidden dimension of 2,048, the required preview window width does not change significantly, which is consistent with the observations in Section 4.1. However, the number of full-view vectors N should increase as the vocabulary size grows. In our experiments, using 5% to 10% of the total vocabulary size as candidates sufficed to achieve a successful approximation. The results prove that the proposed method is scalable and more efficient when applied to large vocabulary softmax.

4.3 Result on machine translation

NMT is based on neural networks and contains an internal softmax function. We applied SVD-softmax to a German to English NMT task to evaluate the actual performance of the proposed algorithm.

The baseline network, which employs the encoder-decoder model with an attention mechanism [25, 26], was trained using the OpenNMT toolkit. The network was trained with concatenated data which contained a WMT 2015 translation task [27], Europarl v7 [28], common crawl [29], and news commentary v10 [30], and evaluated with newstest 2013. The training and evaluation data were tokenized and preprocessed by following the procedures in previous studies [31, 32] to conduct case-sensitive translation with 50,004 frequent words. The baseline network employed 500-dimension word embedding, encoder- and decoder-networks with two unidirectional LSTM layers with 500 units each, and a full-softmax output layer. The network was trained with SGD with an initial learning rate of 1.0 while applying dropout [23] with ratio 0.3 between adjacent LSTM layers. The rest of the training settings followed the OpenNMT training recipe, which is based on