CORPUS	n	# OF TOPICS	PPL	TIME (HOURS)	# OF SERVERS	# OF CLIENTS	# OF TYPES OF $ww_{-n+1}^{-1}g$
44M	3	5	196	0.5	40	100	120.1M
	3	10	194	1.0	40	100	218.6M
	3	20	190	2.7	80	100	537.8M
	3	50	189	6.3	80	100	1.123B
	3	100	189	11.2	80	100	1.616B
	3	200	188	19.3	80	100	2.280B
230M	4	5	146	25.6	280	100	0.681B
1.3B	5	2	111	26.5	400	100	1.790B
	5	5	102	75.0	400	100	4.391B

Table 2: Perplexity (ppl) results and time consumed of composite n-gram/PLSA language model trained on three corpora when different numbers of most likely topics are kept for each document in PLSA.

LANGUAGE MODEL	44M	REDUC-	230M	REDUC-	1.3B	REDUC-
	n=3,m=2	TION	n=4,m=3	TION	n=5,m=4	TION
BASELINE n-GRAM (LINEAR)	262		200		138	
n-GRAM (KNESER-NEY)	244	6.9%	183	8.5%	—	—
m-SLM	279	-6.5%	190	5.0%	137	0.0%
PLSA	825	-214.9%	812	-306.0%	773	-460.0%
n-GRAM+m-SLM	247	5.7%	184	8.0%	129	6.5%
n-GRAM+PLSA	235	10.3%	179	10.5%	128	7.2%
n-GRAM+m-SLM+PLSA	222	15.3%	175	12.5%	123	10.9%
n-GRAM/m-SLM	243	7.3%	171	14.5%	(125)	9.4%
n-GRAM/PLSA	196	25.2%	146	27.0%	102	26.1%
m-SLM/PLSA	198	24.4%	140	30.0%	(103)	25.4%
n-GRAM/PLSA+m-SLM/PLSA	183	30.2%	140	30.0%	(93)	32.6%
n-GRAM/m-SLM+m-SLM/PLSA	183	30.2%	139	30.5%	(94)	31.9%
n-GRAM/m-SLM+n-GRAM/PLSA	184	29.8%	137	31.5%	(91)	34.1%
n-GRAM/m-SLM+n-GRAM/PLSA +m-SLM/PLSA	180	31.3%	130	35.0%	—	—
n-GRAM/m-SLM/PLSA	176	32.8%	—	—	—	—

Table 3: Perplexity results for various language models on test corpus, where + denotes linear combination, / denotes composite model; n denotes the order of n-gram and m denotes the order of SLM; the topic nodes are pruned from 200 to 5.

too big to store in the supercomputer. The composite n-gram/m-SLM/PLSA model gives significant perplexity reductions over baseline n-grams, $n = 3, 4, 5$ and m-SLMs, $m = 2, 3, 4$. The majority of gains comes from PLSA component, but when adding SLM component into n-gram/PLSA, there is a further 10% relative perplexity reduction.

We have applied our composite 5-gram/2-SLM+2-gram/4-SLM+5-gram/PLSA language model that is trained by 1.3 billion word corpus for the task of re-ranking the N-best list in statistical machine translation. We used the same 1000-best list that is used by Zhang et al. (2006). This list was generated on 919 sentences from the MT03 Chinese-English evaluation set by Hiero (Chiang, 2005; Chiang, 2007), a state-of-the-art parsing-based translation model. Its decoder uses a trigram language model trained with modified Kneser-Ney smoothing (Kneser and Ney, 1995) on a 200 million tokens corpus. Each translation has 11 features and language model is one of them. We substitute our language model and use MERT (Och, 2003) to optimize the BLEU score (Papineni et al., 2002). We partition the data into ten pieces, 9 pieces are used as training data to optimize the BLEU score (Papineni et al., 2002) by MERT (Och,