2003), a remaining single piece is used to re-rank the 1000-best list and obtain the BLEU score. The cross-validation process is then repeated 10 times (the folds), with each of the 10 pieces used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation for BLEU score. Table 4 shows the BLEU scores through 10-fold cross-validation. The composite 5-gram/2-SLM+2-gram/4-SLM+5-gram/PLSA language model gives 1.57% BLEU score improvement over the baseline and 0.79% BLEU score improvement over the 5-gram. This is because there is not much diversity on the 1000-best list, and essentially only 20 $\sim$ 30 distinct sentences are there in the 1000-best list. Chiang (2007) studied the performance of machine translation on Hiero, the BLEU score is 33.31% when n-gram is used to re-rank the N-best list, however, the BLEU score becomes significantly higher 37.09% when the n-gram is embedded directly into Hiero's one pass decoder, this is because there is not much diversity in the N-best list. It is expected that putting the our composite language into a one pass decoder of both phrase-based (Koehn et al., 2003) and parsing-based (Chiang, 2005; Chiang, 2007) MT systems should result in much improved BLEU scores.

SYSTEM MODEL	MEAN (%)
BASELINE	31.75
5-GRAM	32.53
5-GRAM/2-SLM+2-GRAM/4-SLM	32.87
5-GRAM/PLSA	33.01
5-GRAM/2-SLM+2-GRAM/4-SLM +5-GRAM/PLSA	33.32

Table 4: 10-fold cross-validation BLEU score results for the task of re-ranking the N-best list.

Besides reporting the BLEU scores, we look at the "readability" of translations similar to the study conducted by Charniak et al. (2003). The translations are sorted into four groups: good/bad syntax crossed with good/bad meaning by human judges, see Table 5. We find that many more sentences are perfect, many more are grammatically correct, and many more are semantically correct. The syntactic language model (Charniak, 2001; Charniak, 2003) only improves translations to have good grammar, but does not improve translations to preserve meaning. The composite 5-gram/2-SLM+2-gram/4-SLM+5-gram/PLSA language model improves both significantly. Bear in mind that Charniak et al. (2003) integrated Charniak's language model with the syntax-based translation model Yamada and Knight proposed (2001) to rescore a tree-to-string translation forest, whereas we use only our language model for N-best list re-ranking. Also, in the same study in (Charniak, 2003), they found that the outputs produced using the n-grams received higher scores from BLEU; ours did not. The difference between human judgments and BLEU scores indicate that closer agreement may be possible by incorporating syntactic structure and semantic information into the BLEU score evaluation. For example, semantically similar words like "insure" and "ensure" in the example of BLEU paper (Papineni et al., 2002) should be substituted in the formula, and there is a weight to measure the goodness of syntactic structure. This modification will lead to a better metric and such information can be provided by our composite language models.

SYSTEM MODEL	P	S	G	W
BASELINE	95	398	20	406
5-GRAM	122	406	24	367
5-GRAM/2-SLM +2-GRAM/4-SLM +5-GRAM/PLSA	151	425	33	310

Table 5: Results of "readability" evaluation on 919 translated sentences, P: perfect, S: only semantically correct, G: only grammatically correct, W: wrong.

5 Conclusion

As far as we know, this is the first work of building a complex large scale distributed language model with a principled approach that is more powerful than n-grams when both trained on a very large corpus with up to a billion tokens. We believe our results still hold on web scale corpora that have trillion tokens, since the composite language model effectively encodes long range dependencies of natural language that n-gram is not viable to consider. Of course, this implies that we have to take a huge amount of resources to perform the computation, nevertheless this becomes feasible, affordable, and cheap in the era of cloud computing.