	MT04 (tune)		MT02		MT03		MT05		Avg
Baseline	18.14		23.87		18.88		22.60		
+POS	18.11	−0.03	23.65	−0.22	18.99	+0.11	22.29	−0.31	−0.17
+POS+Agr	18.86	+**0.72**	24.84	+**0.97**	20.26	+**1.38**	23.48	+**0.88**	+1.04
genres	nw		nw		nw		nw		
#sentences	*1353*		*728*		*663*		*1056*		*2447*

Table 2: Translation quality results (BLEU-4 [%]) for newswire (nw) sets. *Avg* is the weighted averaged (by number of sentences) of the individual test set gains. All improvements are statistically significant at $p \leq 0.01$.

	MT06		MT08		Avg
Baseline	14.68		14.30		
+POS	14.57	−0.11	14.30	+0.0	−0.06
+POS+Agr	15.04	+**0.36**	14.49	+**0.19**	+0.29
genres	nw,bn,ng		nw,ng,wb		
#sentences	*1797*		*1360*		*3157*

Table 3: Mixed genre test set results (BLEU-4 [%]). The MT06 result is statistically significant at $p \leq 0.01$; MT08 is significant at $p \leq 0.02$. The genres are: nw, broadcast news (bn), newsgroups (ng), and weblog (wb).

2008). For each set of results, we initialized MERT with uniform feature weights.

We trained the translation model on 502 million words of parallel text collected from a variety of sources, including the Web. Word alignments were induced using a hidden Markov model based alignment model (Vogel et al., 1996) initialized with bilexical parameters from IBM Model 1 (Brown et al., 1993). Both alignment models were trained using two iterations of the expectation maximization algorithm. Our distributed 4-gram language model was trained on 600 million words of Arabic text, also collected from many sources including the Web (Brants et al., 2007).

For development and evaluation, we used the NIST Arabic-English data sets, each of which contains one set of Arabic sentences and multiple English references. To reverse the translation direction for each data set, we chose the first English reference as the source and the Arabic as the reference.

The NIST sets come in two varieties: newswire (MT02-05) and mixed genre (MT06,08). Newswire contains primarily Modern Standard Arabic (MSA), while the mixed genre data sets also contain transcribed speech and web text. Since the ATB contains MSA, and significant lexical and syntactic differences may exist between MSA and the mixed genres, we achieved best results by tuning on MT04, the largest newswire set.

We evaluated translation quality with BLEU-4 (Papineni et al., 2002) and computed statistical significance with the approximate randomization method of Riezler and Maxwell (2005).[9]

6 Discussion of Translation Results

Tbl. 2 shows translation quality results on newswire, while Tbl. 3 contains results for mixed genres. The baseline is our standard system feature set. For comparison, +POS indicates our class-based model trained on the 11 coarse POS tags only (e.g., “Noun”). Finally, +POS+Agr shows the class-based model with the fine-grained classes (e.g., “Noun+Fem+Sg”).

The best result—a +1.04 BLEU average gain—was achieved when the class-based model training data, MT tuning set, and MT evaluation set contained the same genre. We realized smaller, yet statistically significant, gains on the mixed genre data sets. We tried tuning on both MT06 and MT08, but obtained insignificant gains. In the next section, we investigate this issue further.

Tuning with a Treebank-Trained Feature The class-based model is trained on the ATB, which is predominantly MSA text. This data set is syntactically regular, meaning that it does not have highly dialectal content, foreign scripts, disfluencies, etc. Conversely, the mixed genre data sets contain more irregularities. For example, 57.4% of MT06 comes from non-newswire genres. Of the 764 newsgroup sentences, 112 contain some Latin script tokens, while others contain very little morphology:

[9] With the implementation of Clark et al. (2011), available at: http://github.com/jhclark/multeval.