tune\test	BLEU	TER	TESLA-M	TESLA-F
BLEU	0.5237	0.6029	0.3922	0.4114
TER	**0.5239**	**0.6028**	0.3880	0.4095
TESLA-M	0.5005	0.6359	**0.4170**	0.4223
TESLA-F	0.4992	0.6377	0.4164	**0.4224**

(a) The French-English task

tune\test	BLEU	TER	TESLA-M	TESLA-F
BLEU	0.5641	0.5764	0.4315	0.4328
TER	**0.5667**	**0.5725**	0.4204	0.4282
TESLA-M	0.5253	0.6246	**0.4511**	0.4398
TESLA-F	0.5331	0.6111	0.4498	**0.4409**

(b) The Spanish-English task

tune\test	BLEU	TER	TESLA-M	TESLA-F
BLEU	**0.4963**	**0.6329**	0.3369	0.3927
TER	**0.4963**	0.6355	0.3191	0.3851
TESLA-M	0.4557	0.7055	**0.3784**	**0.4070**
TESLA-F	0.4642	0.6888	0.3753	0.4068

(c) The German-English task

Table 4: Automatic evaluation scores

	P(A)	Kappa
French-English	0.6846	0.5269
Spanish-English	0.6124	0.4185
German-English	0.3973	0.0960

Table 5: Inter-annotator agreement

ric M usually does the best or very close to the best when evaluated by M. On the other hand, the differences between different systems can be substantial, especially between BLEU/TER and TESLA-M/TESLA-F.

In addition to the automatic evaluation, we enlisted twelve judges to manually evaluate the first 200 test sentences. Four judges are assigned to each of the three language pairs. For each test sentence, the judges are presented with the source sentence, the reference English translation, and the output from the four competing Joshua systems. The order of the translation candidates is randomized so that the judges will not see any patterns. The judges are instructed to rank the four candidates, and ties are allowed.

The inter-annotator agreement is reported in Table 5. We consider the judgment for a pair of system outputs as one data point. Let $P(A)$ be the proportion of times that the annotators agree, and $P(E)$

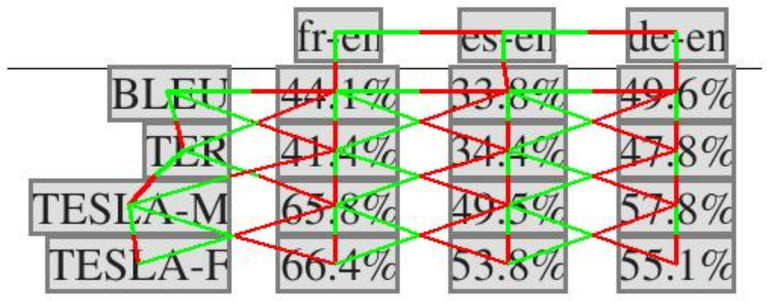

Table 6: Percentage of times each system produces the best translation

be the proportion of times that they would agree by chance. The Kappa coefficient is defined as

$$\text{Kappa} = \frac{\text{P(A)} - \text{P(E)}}{1 - \text{P(E)}}$$

In our experiments, each data point has three possible values: A is preferred, B is preferred, and no preference, hence $P(E) = 1/3$. Our Kappa is calculated in the same way as the WMT workshops (Callison-Burch et al., 2009; Callison-Burch et al., 2010).

Kappa coefficients between 0.4 and 0.6 are considered *moderate*, and our values are in line with those reported in the WMT 2010 translation campaign. The exception is the German-English pair, where the annotators only reach *slight* agreement. This might be caused by the lower quality of German to English translations compared to the other two language pairs.

Table 6 shows the proportion of times each system produces the best translation among the four. We observe that the rankings are largely consistent across different language pairs: Both TESLA-F and TESLA-M strongly outperform BLEU and TER. Note that the values in each column do not add up to 100%, since the candidate translations are often identical, and even a different translation can receive the same human judgment.

Table 7 shows our main result, the pairwise comparison between the four systems for each of the language pairs. Again the rankings consistently show that both TESLA-F and TESLA-M strongly outperform BLEU and TER. All differences are statistically significant under the Sign Test at $p = 0.01$, with the exception of TESLA-M vs TESLA-F in the French-English task, BLEU vs TER in the Spanish-English task, and TESLA-M vs TESLA-F and BLEU vs TER in the German-English task. The results provide strong evidence that tuning machine