

Table 8.3: Baseline and default parameters and methods for all experiments.

Parameter/Method	Baseline setting	Experimental setting	
		Sentence level	System level
Evaluator normalization	none	(0, 1)-normalization	none
Case	ignore case	ignore case	use case
Punctuation	mteval	mteval; treat abbreviations	
Summation of scores	weighted	-	weighted
Reference length	average	best relative sentence	
Sentence boundaries	none	initial and end	
BLEU smoothing	none	BLEU-S	
Substitution cost	constant		
Evaluator aggregation	average		

Table 8.4: Effect of baseline settings and experimental default settings on the correlation with human evaluation. Pearson's r on sentence level.

Hu- man score	Automatic measure + settings	TIDES	TIDES	TIDES	TIDES	TIDES	BTEC	BTEC
		2002 CE	2003 CE	2003 AE	2004 CE	2004 AE	2004 CE	2004 JE
A	WER baseline	0.220	0.256	0.386	0.451	0.542	0.598	0.649
	default	0.320	0.349	0.505	0.540	0.597	0.691	0.744
	PER baseline	0.237	0.313	0.370	0.506	0.538	0.640	0.671
	default	0.329	0.428	0.495	0.579	0.600	0.708	0.744
F	BLEU baseline	0.223	0.284	0.389	0.451	0.503	0.483	0.555
	default	0.404	0.451	0.541	0.606	0.621	0.570	0.635
	NIST baseline	0.388	0.435	0.492	0.563	0.565	0.512	0.577
	default	0.434	0.513	0.562	0.600	0.604	0.520	0.579
A+F	WER baseline	0.178	0.224	0.322	0.438	0.442	0.532	0.582
	default	0.277	0.301	0.423	0.511	0.496	0.565	0.624
	PER baseline	0.170	0.203	0.286	0.435	0.373	0.454	0.495
	default	0.245	0.298	0.389	0.493	0.424	0.456	0.504
A+F	BLEU baseline	0.160	0.193	0.302	0.384	0.391	0.380	0.451
	default	0.354	0.368	0.458	0.540	0.527	0.390	0.462
	NIST baseline	0.280	0.246	0.372	0.428	0.395	0.275	0.339
	default	0.329	0.343	0.440	0.459	0.429	0.277	0.339
A+F	WER baseline	0.220	0.265	0.387	0.476	0.533	0.631	0.683
	default	0.328	0.365	0.518	0.559	0.589	0.702	0.761
	PER baseline	0.227	0.291	0.360	0.507	0.497	0.613	0.650
	default	0.321	0.419	0.496	0.575	0.556	0.653	0.697
A+F	BLEU baseline	0.214	0.268	0.379	0.451	0.485	0.482	0.560
	default	0.416	0.464	0.556	0.612	0.618	0.539	0.612
	NIST baseline	0.372	0.388	0.476	0.537	0.524	0.443	0.513
	default	0.427	0.498	0.563	0.572	0.560	0.448	0.514

Table 8.5: Effect of baseline settings and experimental default settings on the correlation with A + F. Kendall’s τ on sentence level.

Automatic measure + settings	TIDES	TIDES	TIDES	TIDES	TIDES	BTEC	BTEC
	2002 CE	2003 CE	2003 AE	2004 CE	2004 AE	2004 CE	2004 JE
WER baseline default	0.076 0.145	0.126 0.193	0.276 0.372	0.303 0.363	0.290 0.317	0.390 0.389	0.559 0.573
PER baseline default	0.119 0.185	0.173 0.271	0.284 0.366	0.350 0.382	0.291 0.317	0.376 0.364	0.535 0.534
BLEU baseline default	0.121 0.230	0.183 0.286	0.290 0.389	0.322 0.400	0.313 0.328	0.411 0.262	0.537 0.463
NIST baseline default	0.205 0.235	0.234 0.309	0.339 0.397	0.366 0.386	0.302 0.305	0.247 0.248	0.405 0.401

Table 8.6: Effect of baseline settings and experimental default settings on the correlation with A + F. Pearson’s r on system level.

Automatic measure + settings	TIDES	TIDES	TIDES	TIDES	TIDES	BTEC	BTEC
	2002 CE	2003 CE	2003 AE	2004 CE	2004 AE	2004 CE	2004 JE
WER baseline default	-0.056 0.339	0.543 0.813	0.845 0.928	0.918 0.957	0.988 0.994	0.909 0.898	0.949 0.979
PER baseline default	0.064 0.455	0.720 0.907	0.820 0.919	0.967 0.969	0.962 0.965	0.844 0.776	0.933 0.922
BLEU baseline default	0.238 0.618	0.840 0.927	0.925 0.924	0.987 0.989	0.993 0.989	0.890 0.690	0.951 0.923
NIST baseline default	0.436 0.530	0.828 0.907	0.917 0.915	0.952 0.956	0.971 0.985	0.480 0.429	0.782 0.766

Table 8.7: Effect of baseline settings and experimental default settings on the correlation with A + F. Kendall’s τ on system level.

Automatic measure + settings	TIDES	TIDES	TIDES	TIDES	TIDES	BTEC	BTEC
	2002 CE	2003 CE	2003 AE	2004 CE	2004 AE	2004 CE	2004 JE
WER baseline default	0.056 0.167	0.333 0.619	0.600 0.733	0.733 0.822	1.000 1.000	0.745 0.818	0.929 0.929
PER baseline default	0.000 0.278	0.524 0.619	0.467 0.733	0.911 0.822	0.800 0.800	0.636 0.636	0.714 0.714
BLEU baseline default	0.278 0.444	0.619 0.619	0.733 0.733	0.956 0.867	1.000 1.000	0.782 0.564	0.857 0.786
NIST baseline default	0.333 0.389	0.524 0.619	0.733 0.733	0.867 0.778	1.000 1.000	0.455 0.527	0.571 0.571

Table B.1: Effect of baseline settings and experimental default settings on the correlation with human evaluation. Kendall's $\bar{\tau}$ on sentence level.

Hu- man score	Automatic measure + settings	TIDES 2002 CE	TIDES 2003 CE	TIDES 2003 AE	TIDES 2004 CE	TIDES 2004 AE	BTEC 2004 CE	BTEC 2004 JE
A	WER baseline default	0.073 0.141	0.131 0.200	0.291 0.374	0.299 0.351	0.294 0.298	0.378 0.393	0.554 0.584
	PER baseline default	0.117 0.180	0.202 0.276	0.296 0.372	0.360 0.380	0.293 0.315	0.424 0.424	0.587 0.603
	BLEU baseline default	0.115 0.220	0.210 0.296	0.298 0.386	0.325 0.399	0.306 0.325	0.377 0.300	0.532 0.510
	NIST baseline default	0.212 0.233	0.282 0.320	0.355 0.405	0.387 0.396	0.312 0.305	0.311 0.310	0.492 0.487
F	WER baseline default	0.070 0.131	0.106 0.135	0.228 0.284	0.279 0.321	0.250 0.266	0.310 0.312	0.463 0.459
	PER baseline default	0.099 0.158	0.114 0.188	0.238 0.282	0.304 0.313	0.242 0.272	0.272 0.261	0.408 0.392
	BLEU baseline default	0.106 0.208	0.129 0.197	0.249 0.306	0.284 0.325	0.267 0.271	0.366 0.207	0.455 0.355
	NIST baseline default	0.165 0.193	0.145 0.206	0.281 0.314	0.296 0.303	0.248 0.245	0.178 0.179	0.280 0.276
A+F	WER baseline default	0.076 0.145	0.126 0.193	0.276 0.372	0.303 0.363	0.290 0.317	0.390 0.389	0.559 0.573
	PER baseline default	0.119 0.185	0.173 0.271	0.284 0.366	0.350 0.382	0.291 0.317	0.376 0.364	0.535 0.534
	BLEU baseline default	0.121 0.230	0.183 0.286	0.290 0.389	0.322 0.400	0.313 0.328	0.411 0.262	0.537 0.463
	NIST baseline default	0.205 0.235	0.234 0.309	0.339 0.397	0.366 0.386	0.302 0.305	0.247 0.248	0.405 0.401