Song et al. reproduced results using BERT (huggingface):

NOTE: wrong samples per confusion matrix are samples that are tagged with either no BOS tag or more than one BOS tags

Dataset 0_test_bert.txt

		Predicted				
		A	**B**	**C**	**D**	**E**
True	**A**	1033	23	13	33	46
	B	25	1055	10	16	19
	C	19	12	1088	6	11
	D	19	7	14	1092	9
	E	17	15	17	16	1099
Wrong samples		81				

Accuracy: 0.926
F1: 0.939

Dataset 1_test_bert.txt

		Predicted				
		A	**B**	**C**	**D**	**E**
True	**A**	997	16	20	28	30
	B	25	1066	14	19	20
	C	17	15	1103	10	16
	D	24	13	8	1147	10
	E	34	18	11	10	1036
Wrong samples		88				

Accuracy: 0.923
F1: 0.937

Dataset 2_test_bert.txt

Predicted						
True		A	B	C	D	E
	A	1030	26	16	31	46
	B	21	1064	10	7	18
	C	19	9	1134	13	6
	D	15	11	14	1046	11
	E	43	17	13	16	1105
Wrong samples		54				

Accuracy: 0.928
F1: 0.937

Dataset 3_test_bert.txt

Predicted						
True		A	B	C	D	E
	A	1072	17	21	26	29
	B	14	1062	16	13	18
	C	26	13	1111	6	12
	D	27	12	10	1079	12
	E	42	26	13	10	1042
Wrong samples		66				

Accuracy: 0.926
F1: 0.937

Dataset 4_test_bert.txt

Predicted						
True		A	B	C	D	E
	A	1011	13	22	26	35
	B	21	1018	13	14	19
	C	21	11	113	9	11
	D	24	11	16	1102	12
	E	30	25	15	13	1096
Wrong samples		82				

Accuracy: 0.921
F1: 0.934