Song et al. reproduced results using hybrid NER model from deeppavlov:

Model dimensions: **batch size: 50**, **epochs: 50**

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	985	20	29	45	45
	B	24	1033	9	24	23
	C	16	10	1065	16	16
	D	10	9	13	1079	19
	E	15	17	12	17	1067
Wrong samples		177				

Accuracy: 0.902
F1: 0.931

Dataset 1_test_bert.txt

Predicted						
		A	B	C	D	E
True	A	968	21	25	27	28
	B	24	1040	14	20	25
	C	9	14	1072	12	18
	D	25	15	14	1105	16
	E	35	15	10	16	993
Wrong samples		234				

Accuracy: 0.894
F1: 0.931

Dataset 2_test_bert.txt

Predicted						
True		**A**	**B**	**C**	**D**	**E**
	A	976	19	25	35	40
	B	23	1027	10	14	23
	C	13	12	1114	13	7
	D	10	10	23	1026	9
	E	35	15	19	20	1074
Wrong samples		203				

Accuracy: 0.90
F1: 0.933

Dataset 3_test_bert.txt

Predicted						
True		**A**	**B**	**C**	**D**	**E**
	A	1005	25	22	34	42
	B	13	1037	16	16	18
	C	14	17	1095	12	18
	D	13	11	11	1048	23
	E	26	23	14	6	1030
Wrong samples		206				

Accuracy: 0.90
F1: 0.933

Dataset 4_test_bert.txt

Predicted						
True		**A**	**B**	**C**	**D**	**E**
	A	969	27	25	23	36
	B	24	1006	12	14	23
	C	21	20	1093	7	10
	D	20	21	21	1071	15
	E	24	27	22	17	1042
Wrong samples		207				

Accuracy: 0.894
F1: 0.926