Supplementary Material

A Appendix

A.1 Corpus Statistics

Table 1 presents the train/dev/test splits used for the NER model training, along with the total number of tokens present in the training data.

Source	Dataset	Train / Dev / Test # Sentences	Total Tokens in Train
LDC	Hindi	2570 / 809 / 1592	48604
	Indonesian	3181 / 1001 / 1991	55270
	Spanish	1398 / 465 / 928	31799
CoNLL	Dutch	13274 / 2307 / 4227	200059
	German	12067 / 2849 / 2984	206846
	Spanish	8357 / 1915 / 1517	264715

Table 1: Corpus Statistics.

A.2 NER Model Hyperparameters

For each language, we train the model with 100d pre-trained GloVe (Pennington et al., 2014) word embeddings trained on Wikipedia and the monolingual text extracted from the train set. We use hidden size of 200 for each direction of the LSTM and a dropout of 0.5. SGD is used as the optimizer with a learning rate of 0.015. During fine-tuning, the NER model is first trained on the transferred data with the above settings. For the first active learning run, the model is fine-tuned on the target language with a lower learning rate of 1e-5 and for each subsequent run, this rate is increased to 0.015.

A.3 Training Schemes

The results for comparing the different training schemes for Spanish CoNLL, German CoNLL and Indonesian can be seen in Figure 1.

A.4 Variance Analysis

Figure 2 shows the 95% confidence intervals of the NER models comparing the different active learning strategies for the CoNLL datasets using the bootstrap re-sampling method.

Dataset	Tokens	ETAL	SAL	RAND	CFEAL
Dutch CoNLL	200	**69.4 ± 1.6**	**69.6 ± 1.6**	**69.4 ± 1.6**	**69.4 ± 1.6**
	600	74.8 ± 1.6	69.4 ± 1.6	67.2 ± 2.1	66.3 ± 1.8
	1200	77.0 ± 1.5	69.6 ± 1.7	74.0 ± 0.0	68.7 ± 1.8
German CoNLL	200	**59.3 ± 1.7**	**57.4 ± 1.9**	55.2 ± 2.1	54.7 ± 2.1
	600	62.9 ± 1.7	58.7 ± 1.8	58.1 ± 2.0	57.2 ± 1.8
	1200	64.7 ± 1.7	58.7 ± 1.8	60.7 ± 1.8	60.1 ± 1.7
Spanish CoNLL	200	**69.7 ± 1.7**	65.8 ± 1.8	**69.5 ± 1.6**	65.3 ± 1.7
	600	75.3 ± 1.8	66.3 ± 1.8	73.3 ± 1.8	67.8 ± 1.7
	1200	77.1 ± 1.7	65.7 ± 1.8	73.2 ± 1.8	70.2 ± 1.7

Table 2: Variance analysis for significance testing of different active learning systems using paired bootstrap resampling. ± denotes the 95% confidence intervals. Systems which are not statistically significant than the best system ETAL are highlighted in bold.

A.5 Comprehensive Results

Table 3, 4, 5, 5, 7, 8 compares the number of entities present in the data selected by ETAL, CFEAL and SAL across all the datasets.

Tables 9, 10, 11, 12, 13, 14 show the tabulated results for the NER models trained with different active learning strategies for Hindi, Indonesian, German, Spanish and Dutch datasets.

As mentioned in the ablation study which evaluates the effectiveness of PARTIAL-CRF over FULL-CRF, we find that FULL-CRF significantly hurts the recall. Table 15, 16, 17, 18, 19 documents the results of the recall scores across the two settings for Hindi, Indonesian, Spanish-LDC, Spanish-CoNLL, German and Dutch respectively.

References

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

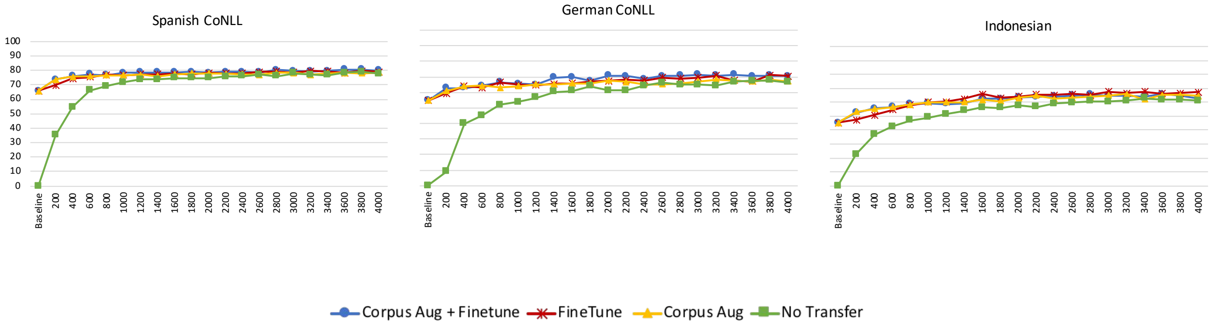

Figure 1: Comparison of the NER performance trained with different schemes for the ETAL strategy.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ETAL + PARTIAL-CRF + CT	115	192	281	379	482	580	675	769	854	934	994	1083	1135	1158	1171	1178	1178	1179	1180	1180
CFEAL+ PARTIAL-CRF + CT	88	207	298	397	506	608	698	793	877	978	1047	1078	1104	1111	1113	1119	1123	1131	1132	1137
SAL+ FULL-CRF + CT	21	42	45	52	60	70	88	95	111	126	133	150	158	174	184	195	210	227	235	246

Table 3: Comparing number of entities across ETAL, SAL and CFEAL for the Hindi LDC dataset.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ETAL + PARTIAL-CRF + CT	87	186	303	413	525	647	741	849	949	1056	1138	1221	1303	1360	1450	1484	1511	1525	1535	1536
CFEAL+ PARTIAL-CRF + CT	86	192	280	371	449	517	601	666	726	793	847	911	973	1021	1069	1125	1186	1244	1269	1329
SAL+ FULL-CRF + CT	7	16	28	39	46	50	63	79	90	106	132	143	158	161	168	187	209	225	231	246

Table 4: Comparing number of entities across ETAL, SAL and CFEAL for the Indonesian LDC dataset.

Method	1	2	3	4	5	6	7	8	9	10	11	12
ETAL + PARTIAL-CRF + CT	84	187	280	391	492	534	585	610	617	619	620	621
CFEAL+ PARTIAL-CRF + CT	79	238	408	530	628	709	777	794	800	801	804	805
SAL+ FULL-CRF + CT	5	10	15	18	20	25	30	46	55	66	80	94

Table 5: Comparing number of entities across ETAL, SAL and CFEAL for the Spanish LDC dataset.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ETAL + PARTIAL-CRF + CT	152	298	427	562	693	823	950	1094	1234	1381	1503	1636	1753	1882	2010	2130	2257	2384	2522	2674
CFEAL+ PARTIAL-CRF + CT	64	128	184	236	293	343	389	440	492	543	593	642	682	729	767	803	873	945	1021	1095
SAL+ FULL-CRF + CT	27	44	66	79	88	102	117	129	132	142	154	172	180	196	223	232	240	252	263	279

Table 6: Comparing number of entities across ETAL, SAL and CFEAL for the Spanish CoNLL dataset.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ETAL + PARTIAL-CRF + CT	154	264	386	513	664	775	883	1016	1153	1275	1365	1490	1588	1730	1827	1954	2064	2121	2211	2329
CFEAL+ PARTIAL-CRF + CT	80	158	217	285	365	424	490	566	640	704	772	847	941	1008	1084	1146	1220	1285	1358	1423
SAL+ FULL-CRF + CT	22	68	74	81	93	101	112	123	135	148	166	175	188	198	205	213	224	230	239	243

Table 7: Comparing number of entities across ETAL, SAL and CFEAL for the German CoNLL dataset.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ETAL + PARTIAL-CRF + CT	166	311	448	584	730	862	1008	1119	1227	1356	1466	1592	1708	1810	1931	2041	2152	2256	2376	2496
CFEAL+ PARTIAL-CRF + CT	89	172	253	342	420	494	581	672	767	855	942	1020	1102	1181	1259	1341	1416	1505	1583	1660
SAL+ FULL-CRF + CT	27	48	69	83	96	107	141	151	160	163	171	188	204	226	237	252	262	275	282	283

Table 8: Comparing number of entities across ETAL, SAL and CFEAL for the Dutch CoNLL dataset.

Type	Method	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Span	ETAL + PARTIAL-CRF + CT	45.0	**54.8**	**60.0**	**64.7**	**68.6**	**69.7**	**70.0**	**71.6**	**72.3**	**73.1**	**74.0**	**73.2**	**73.7**	**74.2**	**75.1**	**74.4**	**74.2**	**73.8**	**74.1**	**73.1**	**74.3**
	ETAL + PARTIAL-CRF	0.0	17.5	30.3	51.3	59.0	61.7	64.8	65.2	66.8	67.7	68.5	68.0	70.1	72.0	72.5	73.0	71.4	72.1	72.2	72.0	72.8
	ETAL + FULL-CRF + CT	45.0	54.2	55.8	57.8	60.0	59.5	61.7	62.0	62.5	63.5	63.7	64.1	64.2	64.3	64.4	65.2	65.1	64.0	64.9	64.9	64.2
Span	CFEAL + PARTIAL-CRF + CT	45.0	47.8	46.7	47.4	47.5	60.0	65.5	66.0	67.3	68.0	68.8	69.2	69.6	69.9	70.6	68.8	70.7	71.4	71.1	71.2	72.1
	RAND + PARTIAL-CRF + CT	45.0	50.2	53.2	56.1	57.4	56.1	56.9	58.2	59.5	59.5	58.9	60.3	61.7	60.7	61.9	62.4	62.2	62.8	63.3	64.5	65.2
Sequence	SAL + FULL-CRF + CT	45.0	49.6	51.2	51.6	52.6	54.4	56.6	58.6	58.8	59.1	61.2	62.2	60.2	60.1	60.4	60.6	62.7	62.9	62.9	63.1	64.2

Table 9: Comparison of NER performance of different active learning strategies for the Hindi LDC dataset. F1 scores are reported. Each column corresponds to NER performance on 200 additional annotated tokens.

Type	Method	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Span	ETAL + PARTIAL-CRF + CT	45.4	47.4	50.8	**54.5**	**58.0**	**60.1**	**60.5**	**62.3**	**65.8**	**63.0**	**64.0**	**65.4**	**65.2**	**65.7**	**65.1**	**67.6**	**66.7**	**67.6**	**66.4**	**66.7**	**67.2**
	ETAL + PARTIAL-CRF	0.0	22.8	36.8	42.6	47.0	49.1	51.5	53.8	56.3	55.9	57.6	56.6	59.1	59.6	60.8	60.4	61.2	62.7	61.7	61.9	60.9
	ETAL + FULL-CRF + CT	45.4	**48.4**	**52.3**	52.4	54.2	54.6	55.2	57.0	57.0	58.4	59.1	59.1	59.5	60.7	60.7	61.3	60.3	60.3	61.2	60.9	60.4
Span	CFEAL + PARTIAL-CRF + CT	45.4	48.5	47.1	46.0	47.5	49.8	49.5	53.9	55.7	54.1	54.9	57.1	55.5	54.7	57.9	56.2	57.9	59.3	58.2	58.6	60.2
	RAND + PARTIAL-CRF + CT	45.4	46.8	48.1	47.2	47.2	51.5	51.9	52.5	52.8	52.4	53.2	53.5	54.6	54.1	56.2	55.2	55.8	56.6	58.4	58.6	56.8
Sequence	SAL + FULL-CRF + CT	45.4	47.9	45.7	44.5	45.1	45.4	44.7	45.4	48.8	47.8	49.2	50.6	50.3	51.8	51.0	49.9	52.0	51.8	52.4	50.4	52.7

Table 10: Comparison of NER performance of different active learning strategies for the Indonesian LDC dataset. F1 scores are reported. Each column corresponds to NER performance on 200 additional annotated tokens.

Type	Method	0	1	2	3	4	5	6	7	8	9	10	11	12
Span	ETAL + PARTIAL-CRF + CT	63.0	**66.3**	**67.9**	65.7	**69.4**	**74.1**	**78.9**	**77.6**	**78.2**	**77.6**	**78.2**	**76.1**	**77.2**
	ETAL + PARTIAL-CRF	0.0	10.9	39.8	58.2	63.3	66.8	70.1	70.3	74.5	72.5	72.3	72.5	71.1
	ETAL + FULL-CRF + CT	63.0	62.9	66.6	**67.2**	68.3	68.0	70.6	70.1	68.5	69.4	69.4	70.6	69.7
Span	CFEAL + PARTIAL-CRF + CT	63.0	62.5	63.9	63.8	64.1	68.2	68.7	67.2	69.3	68.7	71.9	70.2	70.3
	RAND + PARTIAL-CRF + CT	63.0	61.2	61.5	61.9	65.7	65.2	64.6	69.3	67.0	67.3	69.3	69.7	68.7
Sequence	SAL + CT	63.0	62.0	61.8	62.5	61.9	62.3	62.3	62.1	62.3	62.3	62.5	62.7	62.2

Table 11: Comparison of NER performance (F1 scores) of different active learning strategies for the Spanish LDC dataset. Each column, except Run 0, corresponds to NER performance on 200 additional annotated tokens.

Type	Method	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Span	ETAL + PARTIAL-CRF + CT	54.7	59.3	**64.1**	63.0	66.5	65.0	64.7	65.4	66.0	66.8	**67.4**	**67.9**	**67.7**	**69.5**	**69.0**	**69.6**	**70.8**	67.7	67.1	**71.3**	**70.5**
	ETAL + PARTIAL-CRF	0.0	9.0	39.8	45.1	51.9	53.9	56.5	60.3	61.0	64.0	61.2	61.1	64.3	66.0	64.9	65.0	64.4	66.8	67.4	67.6	66.4
	ETAL + FULL-CRF + CT	54.7	**60.7**	63.6	**63.9**	65.4	**66.5**	**66.6**	**66.4**	**67.5**	**66.9**	67.3	66.9	67.7	67.7	68.5	69.3	69.3	**69.8**	**70.7**	71.0	70.2
Span	CFEAL + PARTIAL-CRF + CT	54.7	54.7	55.4	57.2	59.0	61.3	60.2	62.3	62.1	61.4	64.5	63.9	63.5	63.9	65.4	65.0	66.2	65.1	65.8	65.4	66.9
	RAND + PARTIAL-CRF + CT	54.7	55.2	57.0	58.1	59.8	57.7	60.7	59.5	57.4	57.7	59.5	60.5	58.1	59.5	61.0	58.5	58.8	60.2	61.6	61.8	58.7
Sequence	SAL + FULL-CRF + CT	54.7	57.4	57.9	58.8	58.5	59.1	58.7	58.8	58.8	59.5	57.9	57.0	56.6	60.4	60.2	60.5	61.2	60.2	61.8	60.9	60.8

Table 12: Comparison of NER performance of different active learning strategies for the German CoNLL dataset. F1 scores are reported. Each column corresponds to NER performance on 200 annotated tokens.

Type	Method	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Span	ETAL + PARTIAL-CRF + CT	65.7	69.8	**74.4**	**75.3**	**77.0**	**76.5**	**77.1**	**77.4**	**77.7**	**77.2**	**78.4**	**78.0**	**77.9**	**79.0**	**79.3**	**78.7**	**79.5**	**79.1**	**78.3**	**79.7**	**79.0**
	ETAL + PARTIAL-CRF	0.0	36.4	54.0	64.5	70.5	72.9	72.8	73.7	74.3	75.8	75.2	74.1	76.0	76.2	75.7	76.0	76.5	76.8	76.9	77.2	77.8
	ETAL + FULL-CRF + CT	65.7	**72.0**	68.8	71.2	71.7	72.2	72.8	73.3	73.4	72.7	73.3	74.7	74.2	73.9	73.6	74.0	73.9	74.1	74.9	74.5	73.7
Span	CFEAL + PARTIAL-CRF + CT	65.7	65.3	66.9	67.8	70.9	71.0	70.2	71.6	71.2	73.2	73.2	73.2	72.5	72.7	72.6	72.9	72.0	73.6	73.6	73.4	73.8
	RAND + PARTIAL-CRF + CT	65.7	69.5	69.5	70.6	72.1	73.2	70.0	72.0	73.9	73.9	73.6	73.0	71.3	75.7	73.5	74.3	75.1	73.7	74.4	76.2	74.9
Sequence	SAL + FULL-CRF + CT	65.7	65.8	67.4	68.2	68.4	68.2	67.3	67.6	69.4	69.6	69.2	68.9	69.0	69.8	70.0	70.6	71.5	70.7	73.0	70.7	72.7

Table 13: Comparison of NER performance of different active learning strategies for the Spanish CoNLL dataset. F1 scores are reported. Each column corresponds to NER performance on 200 annotated tokens.

Type	Method	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Span	ETAL + PARTIAL-CRF + CT	69.4	69.4	**70.0**	**74.8**	**75.2**	**75.6**	**77.0**	**79.4**	**78.7**	**78.7**	**79.2**	**79.2**	**80.1**	**79.5**	**80.8**	**81.2**	**80.4**	**81.3**	**81.7**	**79.8**	**82.1**
	ETAL + PARTIAL-CRF	0.0	18.1	31.4	47.0	62.9	64.9	67.1	69.3	71.7	72.0	74.7	75.0	73.8	76.3	76.5	75.5	76.5	76.7	77.3	76.5	77.5
	ETAL + FULL-CRF + CT	69.4	**69.6**	69.3	70.4	72.6	72.1	75.7	75.1	75.7	74.8	76.3	76.9	75.4	76.8	75.8	77.0	77.3	76.1	77.2	75.7	76.3
Span	CFEAL + PARTIAL-CRF + CT	69.4	69.5	69.6	69.8	69.6	69.9	69.8	69.7	69.8	69.8	69.8	69.8	69.9	69.6	69.6	69.8	69.6	69.6	69.7	69.7	69.7
	RAND + PARTIAL-CRF + CT	69.4	69.5	69.8	67.2	71.3	72.7	74.0	72.5	72.6	72.7	72.5	73.1	73.9	73.8	73.4	72.8	74.4	74.3	73.1	74.6	74.6
Sequence	SAL + FULL-CRF + CT	69.4	**69.6**	69.7	69.4	69.9	69.8	69.6	69.8	69.9	70.1	69.1	70.3	69.7	69.1	69.9	71.0	68.6	71.9	71.0	71.8	71.4

Table 14: Comparison of NER performance of different active learning strategies for the Dutch CoNLL dataset. F1 scores are reported. Each column corresponds to NER performance on 200 annotated tokens.

Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
ETAL + PARTIAL-CRF + CT	38.6	**51.4**	**56.8**	**59.7**	**60.4**	**61.8**	**63.2**	**64.1**	**65.3**	**66.8**	**68.7**	**65.5**	**66.9**	**67.6**	**69.8**	**69.9**	**71.1**	**68.1**	**68.5**	**68.4**	**70.7**
ETAL + FULL-CRF + CT	38.6	45.8	46.0	48.3	50.6	51.1	52.6	53.0	54.2	55.6	55.9	56.4	56.4	54.6	54.9	56.6	56.2	55.1	57.3	57.5	55.7

Table 15: Comparing recall scores for evaluating the effectiveness of PARTIAL-CRF over FULLCRF for the Hindi LDC dataset.

Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
ETAL + PARTIAL-CRF + CT	51.0	47.1	48.3	52.1	55.0	**57.6**	**61.3**	**59.8**	**64.3**	**61.1**	**63.2**	**64.0**	**64.2**	**64.3**	62.8	**66.6**	**64.5**	**64.6**	**63.0**	**65.3**	**64.2**
ETAL + FULL-CRF + CT	51.0	**51.3**	**55.3**	**54.6**	**56.6**	55.4	56.2	58.6	58.9	60.5	60.8	61.1	61.2	60.7	**63.0**	62.5	60.1	58.9	62.5	62.7	62.3

Table 16: Comparing recall scores for evaluating the effectiveness of PARTIAL-CRF over FULLCRF for the Indonesian LDC dataset.

Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
ETAL + PARTIAL-CRF + CT	57.4	**59.5**	58.5	57.5	**63.9**	**72.2**	**75.1**	**76.0**	**76.0**	**75.5**	**75.6**	**73.7**	**74.6**
ETAL + FULL-CRF + CT	57.4	59.4	**61.7**	**60.6**	61.5	61.9	63.1	62.1	62.1	61.9	63.3	63.0	61.9

Table 17: Comparing recall scores for evaluating the effectiveness of PARTIAL-CRF over FULLCRF for the Spanish LDC dataset.

Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
ETAL + PARTIAL-CRF + CT	45.7	**58.3**	**61.6**	**63.9**	**63.2**	**66.0**	**64.1**	**64.2**	**62.8**	**65.3**	**67.4**	**67.8**	**68.9**	**68.1**	**69.4**	**67.3**	**68.4**	63.5	63.1	**69.5**	65.5
ETAL + FULL-CRF + CT	45.7	52.2	56.6	60.2	61.3	61.3	61.1	62.6	61.1	61.0	61.2	62.8	63.1	63.4	63.8	64.4	65.6	**64.2**	**64.8**	67.2	**65.5**

Table 18: Comparing recall scores for evaluating the effectiveness of PARTIAL-CRF over FULLCRF for the German CoNLL dataset.

Method	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
ETAL + PARTIAL-CRF + CT	65.8	66.4	**70.6**	**73.9**	**75.1**	**75.6**	**76.2**	**79.4**	**78.7**	**78.6**	**79.1**	**78.7**	**79.7**	**79.0**	**80.3**	**80.5**	**79.7**	**81.1**	**81.2**	**78.9**	**81.7**
ETAL + FULL-CRF + CT	65.8	**66.9**	66.1	68.8	70.9	70.8	75.5	74.1	75.4	73.6	75.6	76.5	74.9	76.1	75.3	76.5	77.0	75.1	76.9	75.1	75.5

Table 19: Comparing recall scores for evaluating the effectiveness of PARTIAL-CRF over FULLCRF for the Dutch CoNLL dataset.