Supplementary Material: Multi-Task Learning for Sequence Tagging: An Empirical Study

Soravit Changpinyo, Hexiang Hu, and **Fei Sha**
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
schangpi,hexiangh,feisha@usc.edu

A Comparison between different MTL approaches

Settings	Method	UPOS	XPOS	CHUNK	NER	MWE	SEM	SEMTR	SUPSENSE	COM	FRAME	HYP	Average
STL		95.4	95.04	93.49	88.24	53.07	72.77	74.02	66.81	72.71	62.04	46.73	74.58
(Average) Pairwise	**MULTI-DEC**	94.97	94.65	93.37	87.67	57.21	72.63	74.38	67.39	72.12	61.3	47.99	74.88
	TE⊕DEC	95.0	94.77	93.4	87.72	56.67	72.48	74.25	67.19	71.84	58.65	47.45	74.49
	TE⊕ENC	95.0	94.66	93.32	87.65	55.99	72.49	74.24	67.09	72.09	61.62	47.37	74.68
All	**MULTI-DEC**	95.04	94.31	93.44	86.38	61.43	71.53	74.26	68.1	74.54	59.71	51.41	75.47
	TE⊕DEC	94.95	94.42	93.64	86.8	61.97	71.72	74.36	67.98	74.61	58.14	51.31	75.44
	TE⊕ENC	94.94	94.3	93.7	86.01	59.57	71.58	74.35	68.02	74.61	61.83	49.5	75.31
(Average) All-but-one	**MULTI-DEC**	94.91	94.43	93.65	86.15	61.82	71.09	73.75	68.2	74.42	59.66	50.9	75.36
	TE⊕DEC	94.83	94.4	93.64	86.39	60.55	70.95	73.74	67.81	74.47	58.66	50.86	75.12
	TE⊕ENC	94.77	94.35	93.53	85.96	60.23	70.83	73.64	68.15	74.05	61.23	50.15	75.17

Table 1: Comparison between MTL approaches

In Table 1, we summarize the results of different MTL approaches. We observe no significant differences between those methods.

B Additional results on All-but-one settings

Table 2 and Table 3 compare All and All-but-one settings for **TE⊕DEC** and **TE⊕ENC**, respectively. We show similar results for **MULTI-DEC** in the main text.

	UPOS	XPOS	CHUNK	NER	MWE	SEM	SEMTR	SUPSENSE	COM	FRAME	HYP	#↑	#↓
All	94.95	94.42	93.64	86.8	61.97	71.72	74.36	67.98	74.61	58.14	51.31		
All - UPOS		94.06 ↓	93.44	86.47	60.48	71.08 ↓	73.79	68.1	74.69	58.32	50.83	0	2
All - XPOS	94.38 ↓		93.6	86.68	60.09	70.98 ↓	73.78 ↓	67.9	74.26	58.31	50.6	0	3
All - CHUNK	94.6 ↓	94.29		86.08	60.6	70.39 ↓	73.36 ↓	68.07	74.47	58.73	51.1	0	3
All - NER	94.69 ↓	94.31	93.69		60.48 ↓	70.64 ↓	73.59 ↓	67.51	74.49	58.19	50.44	0	4
All - MWE	94.93	94.46	93.72	86.21 ↓		71.11 ↓	74.04	67.38	74.49	57.6	50.5	0	2
All - SEM	94.86	94.41	93.6	85.97 ↓	59.94 ↓		72.26 ↓	67.35	74.34	59.08	50.48	0	3
All - SEMTR	94.8	94.28	93.56	86.23 ↓	61.23	69.62 ↓		68.16	74.36	58.85	51.5	0	2
All - SUPSENSE	94.82	94.4	93.67	86.49	59.11	71.02 ↓	73.76 ↓		74.69	58.28	51.96	0	2
All - COM	95.19 ↑	94.76 ↑	93.79	86.25 ↓	62.02	72.32	74.92 ↑	67.62		60.72 ↑	50.0 ↓	4	2
All - FRAME	95.03	94.6	93.64	86.68	60.52 ↓	71.11 ↓	73.9	67.69	74.49		51.23	0	2
All - HYP	94.94	94.45	93.69	86.86	61.07	71.22	74.04 ↓	68.32	74.4	58.55		0	1
#↑	1	1	0	0	0	0	1	0	0	1	0		
#↓	3	1	0	4	3	8	6	0	0	0	1		

Table 2: F1 scores for **TE⊕DEC**. We compare All with All-but-one settings (All - ⟨TASK⟩). We test on each task in the columns. Beneficial settings are in green. Harmful setting are in red.

C Detailed results separated by the tasks being tested on

In Table 4-14, we provide F1 scores with standard deviations in all settings. Each table corresponds to a task we test our models on. Rows denote training settings and columns denote MTL approaches.

	UPOS	XPOS	CHUNK	NER	MWE	SEM	SEMTR	SUPSENSE	COM	FRAME	HYP	#↑	#↓
All	94.94	94.3	93.7	86.01	59.57	71.58	74.35	68.02	74.61	61.83	49.5		
All - UPOS		94.0	93.36 ↓	85.98	59.58	70.68	73.66	68.19	74.07	60.51	50.23	0	1
All - XPOS	94.24 ↓		93.29 ↓	85.8	59.81	70.57 ↓	73.64 ↓	68.47	73.94	60.13	50.39	0	4
All - CHUNK	94.66	94.3		85.73	61.58	70.78	73.65	67.87	73.67 ↓	61.73	50.18	0	1
All - NER	94.71	94.25	93.5		59.05	70.58 ↓	73.4 ↓	67.95	74.16	59.96	49.95	0	2
All - MWE	94.94	94.5	93.63	86.1		71.12	73.75	69.0	74.28	61.51	49.81	0	0
All - SEM	94.76	94.32	93.45	85.58	59.47		72.21 ↓	67.77	74.2	61.76	50.15 ↑	1	1
All - SEMTR	94.68	94.25	93.54	86.02	60.59	69.86 ↓		67.96	73.81 ↓	61.31	51.72 ↑	1	2
All - SUPSENSE	94.8	94.27	93.56	86.04	59.25	70.53 ↓	73.27 ↓		74.3	59.98	50.01	0	2
All - COM	95.25 ↑	94.72 ↑	93.82	86.23	60.63	72.38 ↑	75.06 ↑	67.94		63.55	48.77	4	0
All - FRAME	94.84	94.39	93.51 ↓	85.99	61.21	70.78	73.69	68.13	74.3		50.35	0	1
All - HYP	94.86	94.45	93.59	86.1	61.09	71.03 ↓	74.09	68.17	73.78 ↓	61.91		0	2
#↑	1	1	0	0	0	1	1	0	0	0	2		
#↓	1	0	3	0	0	5	4	0	3	0	0		

Table 3: F1 scores for **TE⊕ENC**. We compare All with All-but-one settings (All - ⟨TASK⟩). We test on each task in the columns. Beneficial settings are in green. Harmful setting are in red.

Trained with		Tested on UPOS		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
UPOS only		95.4 ± 0.08		
Pairwise	+XPOS	95.38 ± 0.03	95.4 ± 0.04	95.42 ± 0.07
	+CHUNK	95.43 ± 0.11	95.57 ± 0.02 ↑	95.4 ± 0.0
	+NER	95.38 ± 0.1	95.32 ± 0.03	95.29 ± 0.04
	+MWE	95.15 ± 0.05 ↓	95.11 ± 0.07 ↓	95.05 ± 0.05 ↓
	+SEM	95.23 ± 0.14	95.2 ± 0.05 ↓	95.27 ± 0.08
	+SEMTR	95.17 ± 0.15	95.21 ± 0.03 ↓	95.23 ± 0.13
	+SUPSENSE	95.08 ± 0.08 ↓	95.05 ± 0.04 ↓	95.27 ± 0.08
	+COM	93.04 ± 0.77 ↓	94.03 ± 0.42 ↓	93.6 ± 0.15 ↓
	+FRAME	94.98 ± 0.13 ↓	94.79 ± 0.09 ↓	95.0 ± 0.07 ↓
	+HYP	94.84 ± 0.07 ↓	94.35 ± 0.21 ↓	94.43 ± 0.15 ↓
	Average	94.97	95.0	95.0
All-but-one	All - XPOS	94.57 ± 0.12 ↓	94.38 ± 0.05 ↓	94.24 ± 0.24 ↓
	All - CHUNK	94.84 ± 0.01 ↓	94.6 ± 0.1 ↓	94.66 ± 0.15 ↓
	All - NER	94.81 ± 0.07 ↓	94.69 ± 0.05 ↓	94.71 ± 0.07 ↓
	All - MWE	94.93 ± 0.01 ↓	94.93 ± 0.08 ↓	94.94 ± 0.04 ↓
	All - SEM	94.82 ± 0.17 ↓	94.86 ± 0.08 ↓	94.76 ± 0.15 ↓
	All - SEMTR	94.83 ± 0.12 ↓	94.8 ± 0.03 ↓	94.68 ± 0.17 ↓
	All - SUPSENSE	94.97 ± 0.07 ↓	94.82 ± 0.03 ↓	94.8 ± 0.07 ↓
	All - COM	95.19 ± 0.05 ↓	95.19 ± 0.04 ↓	95.25 ± 0.02 ↓
	All - FRAME	95.15 ± 0.07 ↓	95.03 ± 0.17	94.84 ± 0.1 ↓
	All - HYP	94.93 ± 0.18 ↓	94.94 ± 0.11 ↓	94.86 ± 0.04 ↓
All		95.04 ± 0.03 ↓	94.95 ± 0.08 ↓	94.94 ± 0.1 ↓
Oracle		95.4 ± 0.08	95.57 ± 0.02	95.4 ± 0.08

Table 4: F1 score tested on the task UPOS in different training scenarios

Trained with		Tested on XPOS		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
XPOS only		95.04 ± 0.06		
Pairwise	+UPOS	95.01 ± 0.04	94.99 ± 0.03	94.94 ± 0.05
	+CHUNK	95.1 ± 0.02	95.21 ± 0.02 ↑	95.1 ± 0.04
	+NER	94.98 ± 0.12	95.09 ± 0.07	95.05 ± 0.13
	+MWE	94.7 ± 0.16 ↓	94.8 ± 0.08 ↓	94.66 ± 0.07 ↓
	+SEM	94.77 ± 0.08 ↓	94.82 ± 0.15	94.93 ± 0.08
	+SEMTR	94.86 ± 0.02 ↓	94.8 ± 0.09 ↓	94.97 ± 0.09
	+SUPSENSE	94.75 ± 0.15	94.81 ± 0.06 ↓	95.0 ± 0.12
	+COM	93.19 ± 0.75 ↓	93.94 ± 0.21 ↓	93.12 ± 0.44 ↓
	+FRAME	94.64 ± 0.06 ↓	94.66 ± 0.05 ↓	94.55 ± 0.06 ↓
	+HYP	94.46 ± 0.3 ↓	94.56 ± 0.09 ↓	94.26 ± 0.18 ↓
	Average	94.65	94.77	94.66
All-but-one	All - UPOS	94.03 ± 0.13 ↓	94.06 ± 0.09 ↓	94.0 ± 0.26 ↓
	All - CHUNK	94.46 ± 0.09 ↓	94.29 ± 0.07 ↓	94.3 ± 0.12 ↓
	All - NER	94.3 ± 0.03 ↓	94.31 ± 0.02 ↓	94.25 ± 0.07 ↓
	All - MWE	94.45 ± 0.05 ↓	94.46 ± 0.12 ↓	94.5 ± 0.09 ↓
	All - SEM	94.34 ± 0.09 ↓	94.41 ± 0.09 ↓	94.32 ± 0.17 ↓
	All - SEMTR	94.35 ± 0.08 ↓	94.28 ± 0.07 ↓	94.25 ± 0.12 ↓
	All - SUPSENSE	94.54 ± 0.02 ↓	94.4 ± 0.08 ↓	94.27 ± 0.03 ↓
	All - COM	94.69 ± 0.1 ↓	94.76 ± 0.08 ↓	94.72 ± 0.06 ↓
	All - FRAME	94.57 ± 0.12 ↓	94.6 ± 0.19 ↓	94.39 ± 0.08 ↓
	All - HYP	94.53 ± 0.07 ↓	94.45 ± 0.1 ↓	94.45 ± 0.07 ↓
All		94.31 ± 0.15 ↓	94.42 ± 0.07 ↓	94.3 ± 0.2 ↓
Oracle		95.04 ± 0.06	95.21 ± 0.02	95.04 ± 0.06

Table 5: F1 score tested on the task XPOS in different training scenarios

Trained with		Tested on CHUNK		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
CHUNK only		93.49 ± 0.01		
Pairwise	+UPOS	94.18 ± 0.02 ↑	94.02 ± 0.08 ↑	94.0 ± 0.15 ↑
	+XPOS	93.97 ± 0.16 ↑	94.18 ± 0.01 ↑	93.98 ± 0.13 ↑
	+NER	93.47 ± 0.1	93.64 ± 0.03 ↑	93.54 ± 0.1
	+MWE	93.54 ± 0.13	93.59 ± 0.2	93.33 ± 0.2
	+SEM	93.63 ± 0.02 ↑	93.45 ± 0.07	93.52 ± 0.13
	+SEMTR	93.61 ± 0.07	93.47 ± 0.03	93.45 ± 0.07
	+SUPSENSE	93.2 ± 0.21	93.25 ± 0.15	93.13 ± 0.13 ↓
	+COM	91.94 ± 0.4 ↓	92.29 ± 0.27 ↓	91.86 ± 0.09 ↓
	+FRAME	93.22 ± 0.16 ↓	93.23 ± 0.04 ↓	93.29 ± 0.13
	+HYP	92.96 ± 0.08 ↓	92.86 ± 0.08 ↓	93.13 ± 0.04 ↓
	Average	93.37	93.4	93.32
All-but-one	All - UPOS	93.59 ± 0.13	93.44 ± 0.17	93.36 ± 0.17
	All - XPOS	93.57 ± 0.19	93.6 ± 0.05 ↑	93.29 ± 0.21
	All - NER	93.59 ± 0.09	93.69 ± 0.14	93.5 ± 0.23
	All - MWE	93.71 ± 0.11 ↑	93.72 ± 0.13 ↑	93.63 ± 0.04 ↑
	All - SEM	93.63 ± 0.08	93.6 ± 0.11	93.45 ± 0.13
	All - SEMTR	93.58 ± 0.08	93.56 ± 0.14	93.54 ± 0.06
	All - SUPSENSE	93.67 ± 0.08 ↑	93.67 ± 0.12	93.56 ± 0.12
	All - COM	93.67 ± 0.12	93.79 ± 0.14 ↑	93.82 ± 0.05 ↑
	All - FRAME	93.7 ± 0.09 ↑	93.64 ± 0.11	93.51 ± 0.06
	All - HYP	93.78 ± 0.12 ↑	93.69 ± 0.05 ↑	93.59 ± 0.07
All		93.44 ± 0.09	93.64 ± 0.21	93.7 ± 0.06 ↑
Oracle		94.01 ± 0.13	94.07 ± 0.25	93.93 ± 0.16

Table 6: F1 score tested on the task CHUNK in different training scenarios

Trained with		Tested on NER		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
NER only		88.24 ± 0.09		
Pairwise	+UPOS	87.68 ± 0.41	87.99 ± 0.21	87.43 ± 0.11 ↓
	+XPOS	87.61 ± 0.27 ↓	87.65 ± 0.14 ↓	87.71 ± 0.08 ↓
	+CHUNK	87.96 ± 0.19	88.11 ± 0.21	88.07 ± 0.16
	+MWE	88.15 ± 0.23	87.99 ± 0.15	88.02 ± 0.36
	+SEM	87.35 ± 0.16 ↓	87.27 ± 0.36 ↓	87.49 ± 0.25 ↓
	+SEMTR	87.34 ± 0.27 ↓	87.75 ± 0.38	87.29 ± 0.17 ↓
	+SUPSENSE	87.9 ± 0.24	87.94 ± 0.33	87.92 ± 0.16
	+COM	86.62 ± 0.72 ↓	86.59 ± 0.31 ↓	86.75 ± 0.45 ↓
	+FRAME	88.15 ± 0.35	88.02 ± 0.17	87.99 ± 0.32
	+HYP	87.98 ± 0.21	87.91 ± 0.4	87.82 ± 0.31
	Average	87.67	87.72	87.65
All-but-one	All - UPOS	86.03 ± 0.53 ↓	86.47 ± 0.14 ↓	85.98 ± 0.29 ↓
	All - XPOS	86.04 ± 0.15 ↓	86.68 ± 0.27 ↓	85.8 ± 0.27 ↓
	All - CHUNK	86.05 ± 0.1 ↓	86.08 ± 0.49 ↓	85.73 ± 0.2 ↓
	All - MWE	86.21 ± 0.27 ↓	86.21 ± 0.19 ↓	86.1 ± 0.37 ↓
	All - SEM	85.81 ± 0.32 ↓	85.97 ± 0.14 ↓	85.58 ± 0.04 ↓
	All - SEMTR	86.11 ± 0.28 ↓	86.23 ± 0.23 ↓	86.02 ± 0.39 ↓
	All - SUPSENSE	86.43 ± 0.12 ↓	86.49 ± 0.17 ↓	86.04 ± 0.14 ↓
	All - COM	86.6 ± 0.79 ↓	86.25 ± 0.06 ↓	86.23 ± 0.33 ↓
	All - FRAME	85.9 ± 0.29 ↓	86.68 ± 0.15 ↓	85.99 ± 0.3 ↓
	All - HYP	86.31 ± 0.18 ↓	86.86 ± 0.25 ↓	86.1 ± 0.56 ↓
All		86.38 ± 0.12 ↓	86.8 ± 0.08 ↓	86.01 ± 0.4 ↓
Oracle		88.24 ± 0.09	88.24 ± 0.09	88.24 ± 0.09

Table 7: F1 score tested on the task NER in different training scenarios

Trained with		Tested on MWE		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
MWE only		53.07 ± 0.12		
Pairwise	+UPOS	59.99 ± 0.36 ↑	60.28 ± 0.24 ↑	57.61 ± 0.2 ↑
	+XPOS	58.87 ± 0.78 ↑	60.32 ± 0.3 ↑	58.26 ± 0.25 ↑
	+CHUNK	59.18 ± 0.03 ↑	57.61 ± 1.53 ↑	58.06 ± 0.88 ↑
	+NER	55.4 ± 0.52 ↑	55.17 ± 0.44 ↑	53.4 ± 0.98
	+SEM	60.16 ± 1.23 ↑	58.21 ± 0.09 ↑	58.62 ± 0.61 ↑
	+SEMTR	58.84 ± 1.45 ↑	58.55 ± 0.28 ↑	58.31 ± 2.24 ↑
	+SUPSENSE	58.81 ± 1.01 ↑	58.75 ± 0.33 ↑	58.05 ± 0.72 ↑
	+COM	53.89 ± 1.41	51.72 ± 1.01	51.71 ± 1.05
	+FRAME	53.88 ± 0.76	53.05 ± 1.32	53.3 ± 1.15
	+HYP	53.08 ± 1.72	52.98 ± 1.66	52.59 ± 1.98
	Average	57.21	56.67	55.99
All-but-one	All - UPOS	61.28 ± 0.78 ↑	60.48 ± 0.93 ↑	59.58 ± 1.14 ↑
	All - XPOS	61.91 ± 1.56 ↑	60.09 ± 0.9 ↑	59.81 ± 0.83 ↑
	All - CHUNK	61.01 ± 1.61 ↑	60.6 ± 1.52 ↑	61.58 ± 1.05 ↑
	All - NER	62.69 ± 0.26 ↑	60.48 ± 0.15 ↑	59.05 ± 0.4 ↑
	All - SEM	61.17 ± 0.86 ↑	59.94 ± 0.85 ↑	59.47 ± 0.04 ↑
	All - SEMTR	63.04 ± 0.85 ↑	61.23 ± 2.05 ↑	60.59 ± 0.59 ↑
	All - SUPSENSE	60.51 ± 0.25 ↑	59.11 ± 2.02 ↑	59.25 ± 0.74 ↑
	All - COM	61.95 ± 0.97 ↑	62.02 ± 1.73 ↑	60.63 ± 0.73 ↑
	All - FRAME	62.62 ± 0.85 ↑	60.52 ± 0.47 ↑	61.21 ± 0.99 ↑
	All - HYP	62.04 ± 0.6 ↑	61.07 ± 0.51 ↑	61.09 ± 1.06 ↑
All		61.43 ± 1.94 ↑	61.97 ± 0.5 ↑	59.57 ± 0.64 ↑
Oracle		62.76 ± 0.63	61.74 ± 1.49	61.92 ± 0.66

Table 8: F1 score tested on the task MWE in different training scenarios

Trained with		Tested on SEM		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
SEM only		72.77 ± 0.04		
Pairwise	+UPOS	73.23 ± 0.06 ↑	73.17 ± 0.08 ↑	73.11 ± 0.01 ↑
	+XPOS	73.34 ± 0.12 ↑	73.21 ± 0.04 ↑	73.04 ± 0.21
	+CHUNK	73.16 ± 0.05 ↑	73.02 ± 0.05 ↑	73.13 ± 0.07 ↑
	+NER	72.88 ± 0.08	72.77 ± 0.19	72.91 ± 0.08
	+MWE	72.75 ± 0.09	72.66 ± 0.18	72.83 ± 0.07
	+SEMTR	72.5 ± 0.07 ↓	72.5 ± 0.05 ↓	72.17 ± 0.06 ↓
	+SUPSENSE	72.81 ± 0.04	72.71 ± 0.03	73.09 ± 0.08 ↑
	+COM	70.39 ± 0.46 ↓	70.37 ± 0.28 ↓	70.18 ± 0.54 ↓
	+FRAME	72.76 ± 0.16	72.26 ± 0.21 ↓	72.49 ± 0.23
	+HYP	72.47 ± 0.02 ↓	72.15 ± 0.1 ↓	71.95 ± 1.22
	Average	72.63	72.48	72.49
All-but-one	All - UPOS	70.87 ± 0.19 ↓	71.08 ± 0.19 ↓	70.68 ± 0.76 ↓
	All - XPOS	71.12 ± 0.1 ↓	70.98 ± 0.24 ↓	70.57 ± 0.13 ↓
	All - CHUNK	71.07 ± 0.27 ↓	70.39 ± 0.39 ↓	70.78 ± 0.35 ↓
	All - NER	70.82 ± 0.41 ↓	70.64 ± 0.15 ↓	70.58 ± 0.03 ↓
	All - MWE	71.01 ± 0.14 ↓	71.11 ± 0.17 ↓	71.12 ± 0.29 ↓
	All - SEMTR	69.72 ± 0.27 ↓	69.62 ± 0.37 ↓	69.86 ± 0.36 ↓
	All - SUPSENSE	71.22 ± 0.29 ↓	71.02 ± 0.16 ↓	70.53 ± 0.19 ↓
	All - COM	72.38 ± 0.08 ↓	72.32 ± 0.23 ↓	72.38 ± 0.17 ↓
	All - FRAME	71.48 ± 0.51 ↓	71.11 ± 0.16 ↓	70.78 ± 0.44 ↓
	All - HYP	71.22 ± 0.25 ↓	71.22 ± 0.33 ↓	71.03 ± 0.07 ↓
All		71.53 ± 0.28 ↓	71.72 ± 0.21 ↓	71.58 ± 0.24 ↓
Oracle		73.32 ± 0.04	73.1 ± 0.03	73.14 ± 0.06

Table 9: F1 score tested on the task SEM in different training scenarios

Trained with		Tested on SEMTR		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
SEMTR only		74.02 ± 0.04		
Pairwise	+UPOS	74.93 ± 0.09 ↑	74.87 ± 0.1 ↑	74.85 ± 0.05 ↑
	+XPOS	74.91 ± 0.06 ↑	74.84 ± 0.21 ↑	74.66 ± 0.2 ↑
	+CHUNK	74.79 ± 0.13 ↑	74.73 ± 0.12 ↑	74.77 ± 0.13 ↑
	+NER	74.34 ± 0.08 ↑	74.01 ± 0.05	74.04 ± 0.07
	+MWE	74.51 ± 0.18 ↑	74.63 ± 0.28 ↑	74.66 ± 0.21 ↑
	+SEM	74.73 ± 0.1 ↑	74.72 ± 0.14 ↑	74.41 ± 0.01 ↑
	+SUPSENSE	74.61 ± 0.24 ↑	74.52 ± 0.05 ↑	74.94 ± 0.22 ↑
	+COM	72.6 ± 0.95	71.76 ± 0.88 ↓	71.35 ± 0.95 ↓
	+FRAME	74.18 ± 0.19	74.21 ± 0.37	74.63 ± 0.11 ↑
	+HYP	74.23 ± 0.27	74.19 ± 0.45	74.14 ± 0.23
	Average	74.38	74.25	74.24
All-but-one	All - UPOS	73.54 ± 0.54	73.79 ± 0.46	73.66 ± 0.97
	All - XPOS	74.03 ± 0.11	73.78 ± 0.28	73.64 ± 0.07 ↓
	All - CHUNK	73.97 ± 0.22	73.36 ± 0.05 ↓	73.65 ± 0.39
	All - NER	73.51 ± 0.35	73.59 ± 0.19 ↓	73.4 ± 0.19 ↓
	All - MWE	73.61 ± 0.2 ↓	74.04 ± 0.18	73.75 ± 0.24
	All - SEM	71.97 ± 0.3 ↓	72.26 ± 0.28 ↓	72.21 ± 0.48 ↓
	All - SUPSENSE	73.86 ± 0.09	73.76 ± 0.19	73.27 ± 0.2 ↓
	All - COM	74.75 ± 0.22 ↑	74.92 ± 0.1 ↑	75.06 ± 0.12 ↑
	All - FRAME	74.24 ± 0.37	73.9 ± 0.29	73.69 ± 0.32
	All - HYP	74.02 ± 0.12	74.04 ± 0.17	74.09 ± 0.21
All		74.26 ± 0.1 ↑	74.36 ± 0.03 ↑	74.35 ± 0.29
Oracle		75.23 ± 0.06	75.24 ± 0.13	75.09 ± 0.02

Table 10: F1 score tested on the task SEMTR in different training scenarios

Trained with		Tested on SUPSENSE		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
SUPSENSE only		66.81 ± 0.22		
Pairwise	+UPOS	68.25 ± 0.42 ↑	67.8 ± 0.29 ↑	67.76 ± 0.14 ↑
	+XPOS	67.78 ± 0.4 ↑	68.3 ± 0.71 ↑	67.77 ± 0.15 ↑
	+CHUNK	67.39 ± 0.15 ↑	67.29 ± 0.33	67.36 ± 0.29
	+NER	68.06 ± 0.16 ↑	67.25 ± 0.21	67.57 ± 0.27 ↑
	+MWE	66.88 ± 0.14	66.88 ± 0.24	66.26 ± 0.9
	+SEM	68.29 ± 0.21 ↑	68.46 ± 0.38 ↑	68.1 ± 0.59 ↑
	+SEMTR	68.6 ± 0.81 ↑	68.18 ± 0.39 ↑	67.64 ± 0.92
	+COM	65.57 ± 0.17 ↓	64.98 ± 0.34 ↓	65.55 ± 0.18 ↓
	+FRAME	66.59 ± 0.07	66.2 ± 0.16 ↓	66.75 ± 0.22
	+HYP	66.47 ± 0.24	66.52 ± 0.59	66.16 ± 0.43
	Average	67.39	67.19	67.09
All-but-one	All - UPOS	68.27 ± 0.33 ↑	68.1 ± 0.28 ↑	68.19 ± 0.55 ↑
	All - XPOS	67.99 ± 0.5 ↑	67.9 ± 0.54	68.47 ± 0.18 ↑
	All - CHUNK	68.26 ± 0.48 ↑	68.07 ± 0.28 ↑	67.87 ± 0.32 ↑
	All - NER	68.16 ± 0.26 ↑	67.51 ± 0.4	67.95 ± 0.24 ↑
	All - MWE	68.18 ± 0.62 ↑	67.38 ± 0.22	69.0 ± 0.45 ↑
	All - SEM	67.36 ± 0.42	67.35 ± 0.18	67.77 ± 0.28 ↑
	All - SEMTR	68.17 ± 0.15 ↑	68.16 ± 0.47 ↑	67.96 ± 0.73
	All - COM	68.67 ± 0.37 ↑	67.62 ± 0.6	67.94 ± 0.22 ↑
	All - FRAME	68.47 ± 0.72 ↑	67.69 ± 0.95	68.13 ± 0.39 ↑
	All - HYP	68.46 ± 0.37 ↑	68.32 ± 0.18 ↑	68.17 ± 0.36 ↑
All		68.1 ± 0.54 ↑	67.98 ± 0.29 ↑	68.02 ± 0.21 ↑
Oracle		68.53 ± 0.09	68.22 ± 0.61	69.04 ± 0.44

Table 11: F1 score tested on the task SUPSENSE in different training scenarios

Trained with		Tested on COM		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
COM only		72.71 ± 0.75		
Pairwise	+UPOS	72.46 ± 0.34	72.86 ± 0.12	72.09 ± 0.36
	+XPOS	72.83 ± 0.16	72.87 ± 0.56	72.41 ± 0.51
	+CHUNK	72.44 ± 0.11	73.3 ± 0.15	72.88 ± 0.26
	+NER	70.93 ± 0.73	71.08 ± 0.31 ↓	70.78 ± 0.27 ↓
	+MWE	71.31 ± 0.31	70.93 ± 0.43 ↓	71.36 ± 0.42
	+SEM	72.72 ± 0.22	73.14 ± 0.08	72.25 ± 0.07
	+SEMTR	71.96 ± 0.16	71.74 ± 0.46	72.15 ± 0.5
	+SUPSENSE	72.24 ± 0.27	69.13 ± 0.19 ↓	72.12 ± 0.66
	+FRAME	72.47 ± 0.08	72.89 ± 0.22	72.1 ± 0.93
	+HYP	71.82 ± 0.97	70.47 ± 0.81	72.79 ± 0.97
	Average	72.12	71.84	72.09
All-but-one	All - UPOS	74.42 ± 0.24 ↑	74.69 ± 0.26 ↑	74.07 ± 0.19
	All - XPOS	74.36 ± 0.14 ↑	74.26 ± 0.64	73.94 ± 0.3
	All - CHUNK	74.2 ± 0.13 ↑	74.47 ± 0.26 ↑	73.67 ± 0.23
	All - NER	74.08 ± 0.07 ↑	74.49 ± 0.38 ↑	74.16 ± 0.48
	All - MWE	74.7 ± 0.14 ↑	74.49 ± 0.13 ↑	74.28 ± 0.16 ↑
	All - SEM	74.31 ± 0.1 ↑	74.34 ± 0.42	74.2 ± 0.28
	All - SEMTR	74.2 ± 0.24 ↑	74.36 ± 0.36	73.81 ± 0.16
	All - SUPSENSE	74.24 ± 0.44	74.69 ± 0.52 ↑	74.3 ± 0.13 ↑
	All - FRAME	75.03 ± 0.24 ↑	74.49 ± 0.2 ↑	74.3 ± 0.19 ↑
	All - HYP	74.62 ± 0.14 ↑	74.4 ± 0.06 ↑	73.78 ± 0.05
All		74.54 ± 0.53	74.61 ± 0.24 ↑	74.61 ± 0.32 ↑
Oracle		72.71 ± 0.75	72.71 ± 0.75	72.71 ± 0.75

Table 12: F1 score tested on the task COM in different training scenarios

Trained with		Tested on FRAME		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
FRAME only		62.04 ± 0.74		
Pairwise	+UPOS	62.14 ± 0.35	61.54 ± 0.53	62.27 ± 0.33
	+XPOS	60.77 ± 0.39	61.44 ± 0.06	61.62 ± 1.01
	+CHUNK	62.67 ± 0.47	61.39 ± 0.78	62.98 ± 0.5
	+NER	62.39 ± 0.37	59.25 ± 0.52 ↓	63.02 ± 0.39
	+MWE	61.75 ± 0.21	56.77 ± 2.79	60.61 ± 0.91
	+SEM	61.74 ± 0.27	60.09 ± 0.48 ↓	62.17 ± 0.36
	+SEMTR	62.03 ± 0.41	59.77 ± 0.81	62.79 ± 0.19
	+SUPSENSE	61.94 ± 0.43	55.68 ± 0.61 ↓	61.96 ± 0.18
	+COM	56.52 ± 0.27 ↓	55.25 ± 2.29 ↓	57.65 ± 2.42
	+HYP	61.02 ± 0.62	55.35 ± 0.5 ↓	61.14 ± 1.77
	Average	61.3	58.65	61.62
All-but-one	All - UPOS	58.47 ± 1.0 ↓	58.32 ± 0.35 ↓	60.51 ± 0.1 ↓
	All - XPOS	60.16 ± 0.42 ↓	58.31 ± 0.8 ↓	60.13 ± 1.38
	All - CHUNK	60.01 ± 0.65	58.73 ± 0.68 ↓	61.73 ± 0.48
	All - NER	59.17 ± 0.27 ↓	58.19 ± 0.89 ↓	59.96 ± 0.52 ↓
	All - MWE	59.23 ± 0.33 ↓	57.6 ± 0.82 ↓	61.51 ± 0.43
	All - SEM	58.73 ± 0.67 ↓	59.08 ± 0.84 ↓	61.76 ± 0.52
	All - SEMTR	59.49 ± 0.79 ↓	58.85 ± 0.51 ↓	61.31 ± 1.16
	All - SUPSENSE	59.23 ± 0.64 ↓	58.28 ± 0.19 ↓	59.98 ± 1.23
	All - COM	62.37 ± 0.37	60.72 ± 0.73	63.55 ± 0.31
	All - HYP	59.69 ± 0.41 ↓	58.55 ± 0.29 ↓	61.91 ± 0.59
All		59.71 ± 0.85	58.14 ± 0.23 ↓	61.83 ± 0.98
Oracle		62.04 ± 0.74	62.04 ± 0.74	62.04 ± 0.74

Table 13: F1 score tested on the task FRAME in different training scenarios

Trained with		Tested on HYP		
		MULTI-DEC	TE⊕DEC	TE⊕ENC
HYP only		46.73 ± 0.55		
Pairwise	+UPOS	48.02 ± 0.31	49.36 ± 0.36 ↑	48.27 ± 0.68
	+XPOS	48.81 ± 0.36 ↑	49.23 ± 0.55 ↑	48.06 ± 0.02 ↑
	+CHUNK	47.85 ± 0.2 ↑	48.43 ± 0.3 ↑	47.13 ± 0.35
	+NER	47.9 ± 0.67	48.24 ± 0.65	48.64 ± 1.17
	+MWE	47.32 ± 0.29	45.83 ± 0.46	46.71 ± 0.64
	+SEM	48.15 ± 0.21 ↑	47.95 ± 0.75	47.12 ± 0.43
	+SEMTR	47.74 ± 0.57	46.96 ± 0.85	46.1 ± 0.11
	+SUPSENSE	49.23 ± 0.13 ↑	47.29 ± 0.41	47.24 ± 0.43
	+COM	47.41 ± 1.18	45.24 ± 0.46	47.81 ± 0.8
	+FRAME	47.5 ± 0.46	46.0 ± 0.53	46.66 ± 0.54
	Average	47.99	47.45	47.37
All-but-one	All - UPOS	51.13 ± 0.94 ↑	50.83 ± 0.65 ↑	50.23 ± 0.73 ↑
	All - XPOS	51.65 ± 0.63 ↑	50.6 ± 0.44 ↑	50.39 ± 1.17 ↑
	All - CHUNK	50.27 ± 0.76 ↑	51.1 ± 0.28 ↑	50.18 ± 0.81 ↑
	All - NER	50.86 ± 0.87 ↑	50.44 ± 0.39 ↑	49.95 ± 0.38 ↑
	All - MWE	50.83 ± 0.61 ↑	50.5 ± 0.9 ↑	49.81 ± 0.44 ↑
	All - SEM	50.93 ± 0.27 ↑	50.48 ± 0.53 ↑	50.15 ± 0.11 ↑
	All - SEMTR	51.27 ± 0.5 ↑	51.5 ± 0.46 ↑	51.72 ± 0.15 ↑
	All - SUPSENSE	50.86 ± 1.85 ↑	51.96 ± 0.29 ↑	50.01 ± 1.13 ↑
	All - COM	50.28 ± 1.02 ↑	50.0 ± 0.11 ↑	48.77 ± 0.54 ↑
	All - FRAME	50.89 ± 0.64 ↑	51.23 ± 1.01 ↑	50.35 ± 0.68 ↑
All		51.41 ± 0.25 ↑	51.31 ± 0.55 ↑	49.5 ± 0.05 ↑
Oracle		50.0 ± 0.42	50.15 ± 0.25	48.06 ± 0.02

Table 14: F1 score tested on the task HYP in different training scenarios