B Experiment result on full ChEMU-Ref dataset

Below are the experiment results on the full ChEMU-Ref dataset - train/dev/test = 900/225/375 snippets.

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
Coref. (Surface)	coreference	89.4	55.9	68.7	79.2	47.7	59.5
	joint_train	91.4	56.0	69.5	81.3	48.0	60.3
Coref. (Atom)	coreference	89.4	55.9	68.7	81.3	48.3	60.6
	joint_train	91.4	56.0	69.5	83.9	48.8	61.7
Bridging	bridging	89.5	83.9	86.6	81.4	72.8	76.8
	joint_train	91.2	84.1	87.5	83.1	74.1	78.3
TR	bridging	78.6	84.7	81.5	77.4	84.7	80.8
	joint_train	79.7	85.9	82.7	77.6	85.9	81.5
RA	bridging	89.5	84.6	87.0	80.6	68.5	74.0
	joint_train	91.4	85.6	88.4	82.7	69.2	75.3
WU	bridging	91.5	84.0	87.5	81.9	74.3	77.9
	joint_train	93.1	83.7	88.1	83.6	76.0	79.6
CT	bridging	89.8	77.5	83.1	85.1	70.0	76.8
	joint_train	91.3	77.0	83.3	85.9	69.4	76.4
Overall	joint_train	91.2	74.0	81.7	82.8	68.7	75.1

Table 13: Anaphora resolution results over the test dataset (%). Models are trained for "coreference", "bridging" or "joint_train" (both tasks jointly). Models were trained over 30,000 epochs, and averaged over 3 runs with different random seeds. "F_A" and "F_R" denote the F1 score for anaphor and relation prediction, respectively.

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
Coref. (Surface)	coreference	89.4	55.9	68.7	79.2	47.7	59.5
	- w/ oracle mentions	98.1	73.0	83.7	95.0	68.0	79.3
	joint_train	91.4	56.0	69.5	81.3	48.0	60.3
	- w/ oracle mentions	97.3	69.4	81.0	93.3	64.1	76.0
Coref. (Atom)	coreference	89.4	55.9	68.7	81.3	48.3	60.6
	- w/ oracle mentions	98.1	73.0	83.7	96.4	68.4	80.0
	joint_train	91.4	56.0	69.5	83.9	48.8	61.7
	- w/ oracle mentions	97.3	69.4	81.0	95.0	64.5	76.8
Bridging	bridging	89.5	83.9	86.6	81.4	72.8	76.8
	- w/ oracle mentions	94.8	90.9	92.8	94.1	84.8	89.2
TR	joint_train	91.2	84.1	87.5	83.1	74.1	78.3
	- w/ oracle mentions	95.0	90.3	92.6	93.1	83.3	88.0
	bridging	78.6	84.7	81.5	77.4	84.7	80.8
	- w/ oracle mentions	85.6	91.0	88.2	85.6	91.0	88.2
RA	joint_train	79.7	85.9	82.7	77.6	85.9	81.5
	- w/ oracle mentions	88.3	91.6	89.8	86.9	91.2	88.9
	bridging	89.5	84.6	87.0	80.6	68.5	74.0
	- w/ oracle mentions	91.9	92.7	92.2	91.6	81.5	86.2
WU	joint_train	91.4	85.6	88.4	82.7	69.2	75.3
	- w/ oracle mentions	93.5	91.4	92.4	90.6	79.1	84.5
	bridging	91.5	84.0	87.5	81.9	74.3	77.9
	- w/ oracle mentions	97.3	89.8	93.4	95.8	85.7	90.4
CT	joint_train	93.1	83.7	88.1	83.6	76.0	79.6
	- w/ oracle mentions	96.9	89.3	93.0	94.9	84.6	89.4
	bridging	89.8	77.5	83.1	85.1	70.0	76.8
	- w/ oracle mentions	99.1	95.5	97.2	98.8	90.1	94.2
	joint_train	91.3	77.0	83.3	85.9	69.4	76.4
	- w/ oracle mentions	93.9	95.0	94.4	90.9	89.0	89.9
Overall	joint_train	91.2	74.0	81.7	82.8	68.7	75.1
	- w/ oracle mentions	95.7	82.8	88.8	93.1	79.3	85.7

Table 14: Test results with gold-standard mentions during training. Models trained for "coreference", "bridging" or "joint_train" (both tasks jointly). Models trained over 30,000 epochs; averaged over 3 runs with different random seeds. "F_A" and "F_R" denote the F1 score for anaphor and relation prediction, respectively.

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
Coref. (Surface)	coreference	89.4	55.9	68.7	79.2	47.7	59.5
	- w/ CHELMO	91.5	56.6	70.0	83.7	49.8	62.5
Coref. (Atom)	joint_train	91.4	56.0	69.5	81.3	48.0	60.3
	- w/ CHELMO	90.8	57.4	70.3	82.8	50.3	62.6
	coreference	89.4	55.9	68.7	81.3	48.3	60.6
	- w/ CHELMO	91.5	56.6	70.0	86.0	50.5	63.6
	joint_train	91.4	56.0	69.5	83.9	48.8	61.7
	- w/ CHELMO	90.8	57.4	70.3	85.5	51.1	64.0
Bridging	bridging	89.5	83.9	86.6	81.4	72.8	76.8
	- w/ CHELMO	92.8	84.4	88.4	86.3	74.8	80.2
TR	joint_train	91.2	84.1	87.5	83.1	74.1	78.3
	- w/ CHELMO	92.3	85.6	88.8	86.2	75.9	80.7
	bridging	78.6	84.7	81.5	77.4	84.7	80.8
	- w/ CHELMO	80.5	86.9	83.6	78.3	86.9	82.4
RA	joint_train	79.7	85.9	82.7	77.6	85.9	81.5
	- w/ CHELMO	81.9	86.1	83.9	80.5	86.1	83.2
	bridging	89.5	84.6	87.0	80.6	68.5	74.0
	- w/ CHELMO	93.0	85.9	89.3	85.8	70.4	77.3
WU	joint_train	91.4	85.6	88.4	82.7	69.2	75.3
	- w/ CHELMO	92.8	88.0	90.3	84.7	72.7	78.2
	bridging	91.5	84.0	87.5	81.9	74.3	77.9
	- w/ CHELMO	94.7	84.0	89.0	87.1	76.4	81.4
CT	joint_train	93.1	83.7	88.1	83.6	76.0	79.6
	- w/ CHELMO	93.9	84.9	89.2	87.5	77.0	81.9
	bridging	89.8	77.5	83.1	85.1	70.0	76.8
	- w/ CHELMO	96.1	75.7	84.6	90.1	71.4	79.6
	joint_train	91.3	77.0	83.3	85.9	69.4	76.4
	- w/ CHELMO	92.1	80.6	85.7	85.0	73.6	78.6
Overall	joint_train	91.2	74.0	81.7	82.8	68.7	75.1
	- w/ CHELMO	91.9	75.5	82.9	85.7	70.6	77.4

Table 15: Results with different pretrained embeddings. "coreference", "bridging" and "joint_training" represent models that are trained on the coreference resolution task, bridging task, and both tasks jointly, respectively. We train the models over 30,000 epochs, and averages over 3 runs with different random seeds. "F_A" and "F_R" denote the F1 score for anaphor and relation prediction, respectively.