Label Confidence Weighted Learning for Target-Level Sentence Simplification

Xinying Qiu and **Jingshen Zhang**

Department of Computer Science, School of Information Science and Technology

Guangdong University of Foreign Studies, Guangzhou, China

: https://arxiv.org/abs/2410.05748 : xy.qiu@foxmail.com; audbut0702@163.com

EMNLP 2024

Abstract

- **Definition**

Multi-level sentence simplification generates simplified sentences for specific target audiences with varying language proficiency .

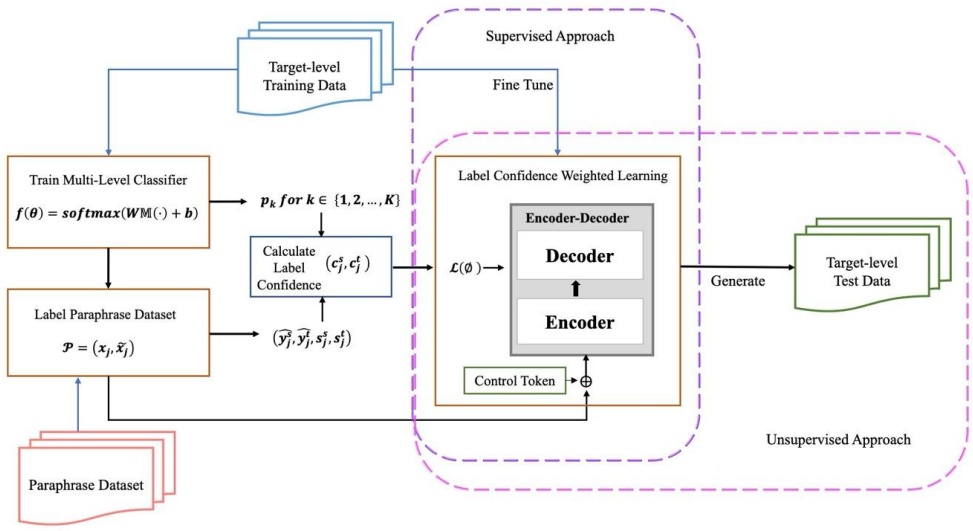
- **Challenge**

Progress in this area has been hindered by limited availability of labeled parallel corpora and the augmentation techniques proposed in previous studies suffer from the propagation of label errors.

- **Proposal**

We propose *Label Confidence Weighted Learning* (LCWL), a novel approach that incorporates a label confidence weighting scheme in the training loss of the encoder-decoder model.

Research Structure



Research Structure with Label Confidence Weighted Learning

Experiment Design

- **Dataset**

Newsela-auto, Para-NMT-50M

- **Baseline Models**

Unsupervised: MUSS, FUDGE-Target, SCE, LCWL, LCWL+SCE

Supervised: SUPER, GPT-3.5-Turbo, SCE+FT, LCWL+FT, SCE+LCWL+FT

- **Evaluation Metrics**

ΔSLE, SARI, FKGL, LENS, LENS-SALSA, BERTScore, BLEU

Results and Analysis

	Evaluation Metrics	*ΔSLE*	LENS↑	LENS-SALSA	*SARI* ↑	FKGL	BS↑	BLEU↑	Avg. Rank↓
Simp-1		0.67		70.697	-	9.66	-	-	
Unsupervised Methods	MUSS	0.69*	**63.28**	**70.46**	35.69	7.75	75.95	41.29	3.15‖**2.4**
	FUDGE	0.32	61.13*	66.73	36.1	**8.81**	80.45	**51.98**	2.86‖3.2
	SCE	0.479	59.28	68.88	37.06*	11.45*	78.1	41.8	3.43‖3.2
	LCWL	1.73	60.86	69.43*	**37.78**	7.1	87.41*	43.07	2.86‖3
	LCWL+SCE	**0.671**	59.68	69.05	37.03	11.92	**88.11**	43.74*	**2.57**‖3
Supervised Methods	SUPER	0.07	65.03*	66.981	32.5	**9.36**	88.2	**75.06**	3.3‖3.6
	GPT-3.5-Turbo	**0.75**	64.76	72.13	**38.45**	10.56	86.29	36.27	3‖**2.2**
	SCE+FT	0.28*	61.59	**69.77**	37.53*	10.57	88.75*	58.94*	**2.7**‖3
	SCE+LCWL+FT	0.277	**65.43**	67.91	36.8	10.51	88.3	58.8	3‖3
	LCWL **+FT**	0.19	63.19	68.28*	37.1	10.29*	**90.08**	57.3	3‖3.2
Simp-2		1.3		73.557	-	7.48	-	-	
Unsupervised Methods	MUSS	0.77	60.27*	**71.3**	36.57	7.27*	65.91	17.23	3.29‖2.6
	FUDGE	0.51	58.19	67.08	**38.32**	**7.42**	70.75	36.89	3.57‖3.4
	SCE	0.595	58.91	69.68	37.33	10.58	**89.61**	37.46	3.43‖4
	LCWL	**1.74**	**61.84**	70.78*	38.27*	7.14	89.54*	38.15*	**1.86‖1.8**
	LCWL+SCE	0.79*	59.99	70.36	37.75	10.83	87.16	**38.71**	2.86‖3.2
Supervised Methods	SUPER	0.14	62.2	66.7	31.1	8.88	78.2	**56.65**	4.3‖4.8
	GPT-3.5-Turbo	**0.88**	**67.16**	**74.02**	41.62*	9.58	87.8	33.4	2.7‖**2**
	SCE+FT	0.61	63.8	72.04*	39.4	**7.82**	**96.92**	52.3	2.7‖3
	SCE+*LCWL*+FT	0.73	65.26*	71	**42.7**	8.35	**96.92**	55.58*	**2.3**‖2.6
	LCWL **+FT**	0.8*	64.7	71.9	41.6	8.3*	**96.92**	48.4	2.6‖2.6
Simp-3		2.0		75.669	-	5.88	-	-	
Unsupervised Methods	MUSS	1.49*	57.02	71.3*	38.05	5.19*	56.03	10.55	3.43‖2.8
	FUDGE	0.81	52.69	68.23	**39.56**	**6.44**	61.46	23.98	3.43‖3.2
	SCE	0.65	58.17	70.68	37.51	10	**89.61**	33.19	3.57‖4.2
	LCWL	**1.63**	**61.68**	**71.83**	38.68*	7.39	89.54*	33.31*	**1.71‖1.6**
	LCWL+SCE	0.83	59.51*	71.08	38.22	10.13	87.16	**34.09**	2.86‖3.2
Supervised Methods	SUPER	0.66	61	66.5	37.9	6.65	66.6	39.6	4.4‖4.6
	GPT-3.5-Turbo	0.97	66.2	74.4	41	8.81	87.79*	31.3	4‖4.2
	SCE+FT	1.59	66.64*	74.4	41.7	5.4	82.9	40.29*	2.9‖3
	SCE+LCWL+FT	1.81*	**67.98**	**75.13**	46.14*	6.1*	**92.15**	**46.48**	**1.4‖1.6**
	LCWL **+FT**	**2.19**	64.8	76.85*	**47.11**	**5.87**	74.7	34	2.3‖**1.6**
Simp-4		2.63		77.639	-	4.16	-	-	
Unsupervised Methods	MUSS	1.41*	55.23	71.15	**39.63**	5.61*	51.73	7.65	3.14‖2.6
	FUDGE	1.04	41.64	61.69	37.03	**4.6**	49.6	11.06	3.86‖3.6
	SCE	0.69	58.94	71.16	35.18	8.81	**87.77**	26.9	3.43‖4
	LCWL	**1.76**	**60.46**	**72.14**	37.49*	5.65	83.72*	**27.48**	**1.57‖1.6**
	LCWL+SCE	0.852	59.89*	71.71*	37.32	9.32	82.85	27.07*	3‖3.2
Supervised Methods	SUPER	1.53	58.9	62.6	43.2	5.09	55	24.5	4.3‖4
	GPT-3.5-Turbo	1.14	65.6	75.1	40.9	7.87	**79.97**	28.6	3.9‖4.6
	SCE+FT	1.98	**67.16**	74.7	42.1	3.95*	63.7	31.47*	2.9‖2.8
	SCE+LCWL+FT	2.33*	62.9	76.52*	**46.42**	**4.01**	65	**39.27**	**1.9‖1.8**
	LCWL **+FT**	**2.69**	65.42*	**77.52**	46.23*	4.73	75.51*	26.4	2.1‖**1.8**

Table 2: Comparison of unsupervised and supervised methods on Newsela-auto across 4 simplification levels using 7 evaluation metrics. ↑ indicates higher scores are better. For ΔSLE, LENS-SALSA, and FKGL, scores closer to the ground truth are better. The best and second-best performances are bolded and starred, respectively. Our proposed methodologies are bolded and italicized. Average ranks within supervised and unsupervised categories are calculated across 7 metrics and 5 metrics (excluding BERTScore and BLUE), separated by ‖ with the 7-metric average on the left, with highest ranks bolded.

Avg. Rank ↓ over 4 Levels	7-Metric	5-Metric	*ΔSLE*	LENS	LENS-SALSA	SARI	FKGL
SCE+LCWL+FT	**1.68**	1.8	2	**1.25**	2	**1.75**	2
LCWL+FT	1.71	**1.6**	**1.5**	1.75	**1.5**	**1.75**	**1.5**
LCWL	2.57	2.6	2.5	3	2.5	2.5	2.5

Table 5: Comparison of average ranks of **two best supervised methods** (SCE+LCWL+FT and LCWL+FT) and **the best unsupervised method** (LCWL). The best ranks are bolded.

Conclusion

- LCWL leverages weak supervision from a large paraphrase dataset and a pre-trained classifier, which sets it apart from existing confidence-weighting methods that primarily focus on classification tasks.
- Experiments on the Newsela-auto dataset demonstrated that LCWL outperforms state-of-the-art unsupervised baselines. After fine-tuning on in-domain labeled data, it consistently delivers superior simplifications compared to strong supervised methods.