Dynamic Multi-Level Multi-Task Learning for Sentence Simplification

Han Guo, Ramakanth Pasunuru, and Mohit Bansal
www.han-guo.info, www.rama-kanth.com, www.cs.unc.edu/~mbansal

Abstract

Sentence simplification aims to improve readability and understandability, based on several operations such as splitting, deletion, and paraphrasing. We have following contributions in this work:

1. We first present a **strong pointer-copy mechanism** based sequence-to-sequence sentence simplification model.
2. We then improve its entailment and paraphrasing capabilities via **multi-task learning with related auxiliary tasks** of entailment and paraphrase generation.
3. We propose a novel **"multi-level" layered soft sharing approach** where each auxiliary task shares different (higher versus lower) level layers with the sentence simplification model, depending on the task's semantic versus lexico-syntactic nature.
4. We also introduce a novel **"dynamic mixing ratio" multi-armed bandit based training approach** that continually learns how to effectively switch across tasks during multi-task learning.

Auxiliary Tasks

- **Entailment generation** is the task of generating a hypothesis which is entailed by the given input premise.
  - A good simplified sentence should be entailed by the source sentence, and hence we incorporate such knowledge through an entailment generation task into our sentence simplification task.
  - We share the higher-level semantic layers between the two tasks.
  - We use entailment pairs from SNLI (Bowman et al., 2015) and Multi-NLI (Williams et al., 2017) datasets for training our entailment generation model.
- **Paraphrase generation** is the task of generating similar meaning phrases or sentences by reordering and modifying the lexicon and/or syntax.
  - Paraphrasing is one of the common operations used in sentence simplification, i.e, by substituting complex words and phrases with their simpler paraphrase forms.
  - We add this knowledge to the sentence simplification task via multi-task learning, by sharing the lower-level lexico-syntactic layers between the two tasks.
  - We use the paraphrase pairs from ParaNMT (Wieting and Gimpel, 2017).

Models

Pointer-Copy Model

- Our sentence simplification base model is a sequence-to-sequence two-layer bidirectional encoder and unidirectional decoder LSTM-RNN, with attention (Bahdanau et al., 2015) and pointer-copy (See et al., 2017).

$$P_f(y) = p_g \cdot P_v(y) + (1 - p_g) \cdot P_c(y)$$

Soft Sharing

- In multi-task learning, we can do either hard sharing or soft sharing of parameters. Hard sharing directly ties the parameters to be shared. Soft sharing only loosely couples the parameters and encourages them to be close in representation space.

$$L(\theta) = -\log P_f(y|x;\theta) + \lambda||\theta_s - \phi_s||$$

Models

Multi-Level Sharing Mechanism

- Recently, Belinkov et al (2017) observed that different layers in a sequence-to-sequence model (trained on translation) exhibit different functionalities: lower-layers (closer to inputs) of the encoder learn to represent word structure while higher layers (farther from inputs) are more focused on semantics and meanings.

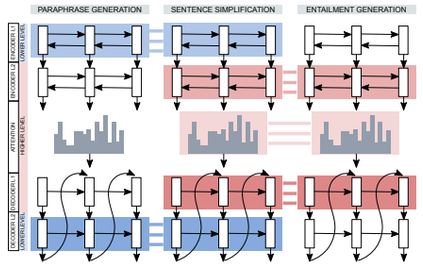

Figure 1: Overview of our 3-way multi-task model. Same color and dashed connections represent soft-shared parameters in different layers.

- Based on these findings, we share the higher-level layers between the entailment generation and sentence simplification tasks, since they share higher semantic-level (full-sentence) language inference skills.
- On the other hand, we share the lower-level lexico-syntactic layers between the paraphrase generation and sentence simplification tasks, since they share more word/phrase and syntactic level paraphrasing knowledge to simplify the smaller, intermediate sub-sentence pieces.

Dynamic Mixing Ratio Learning

- Current multi-task models are trained via alternate mini-batch optimization based on a task "mixing ratio". This is usually treated as a very important hyperparameter to be tuned, but the search space scales exponentially with the number of tasks. Hence, we replace this manually-tuned and static mixing ratio with a "dynamic" mixing ratio learning approach.
- We view the problem of learning the right mixing of tasks as a sequential control problem, where the controller's goal is to decide the next task/action after every certain training steps in each task-sampling round.

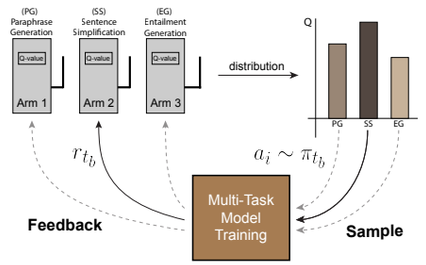

Figure 2: Our multi-armed bandits algorithm for dynamic mixing ratio learning, consisting of a 3-arms/tasks controller.

- We model the controller as an M-armed bandit, where it selects a sequence of actions/arms over the current training trajectory to maximize the expected future payoffs.

$$\pi_{t_b}(a_i) = \exp(Q_{t_b,i}/\tau) \Big/ \sum_{j=1}^{M} \exp(Q_{t_b,j}/\tau) \qquad Q_{t_b,i} = (1-\alpha)^{t_b} Q_{0,i} + \sum_{k=1}^{t_b} \alpha(1-\alpha)^{t_b-k} r_k$$

- One problem in bandits learning is the trade-off between exploration and exploitation. For this, we use the Boltzmann exploration with exponentially moving action value estimates. To further help the exploration process, we follow the principle of optimism under uncertainty and set initial Q to be above the maximum empirical rewards.
- Previous works (Graves et al. 2017, Sharma and Ravindran, 2017) use MAB for choosing domain/data type or NN syllabus for learning efficiency.

Results and Analysis

Setup: We use 3 simplification datasets: Newsela, WikiLarge, WikiSmall

Models	BLEU	FKGL	SARI
PREVIOUS WORK			
PBMT-R	18.19	7.59	15.77
Hybrid	14.46	4.01	30.00
EncDecA	21.70	5.11	24.12
DRESS	23.21	4.13	27.37
DRESS-LS	24.30	4.21	26.63
OUR MODELS			
Baseline ⊗	23.72	3.25	28.31
⊗ + Ent.	16.82	2.21	31.55
⊗ + Paraphr.	16.29	2.03	31.71
⊗+Ent+Par	11.86	1.38	32.98

Table 1: Newsela (FKGL: lower is better). Note that SARI is the primary, human-correlated metric for sentence simplification (Xu et al., 2016).

Models	WIKISMALL BLEU	WIKISMALL FKGL	WIKISMALL SARI	WIKILARGE BLEU	WIKILARGE FKGL	WIKILARGE SARI
PREVIOUS WORK						
PBMT-R	46.31	11.42	15.97	81.11	8.33	38.56
Hybrid	53.94	9.21	30.46	48.97	4.56	31.40
SBMT-SARI	-	-	-	73.08	7.29	39.96
EncDecA	47.93	11.35	13.61	88.85	8.41	35.66
DRESS	34.53	7.48	27.48	77.18	6.58	37.08
DRESS-LS	36.32	7.55	27.24	80.12	6.62	37.27
OUR MODELS						
Baseline ⊗	36.18	7.69	25.67	82.37	7.84	36.68
⊗+Ent+Par	29.70	6.93	28.24	81.49	7.41	37.45

Table 2: WikiSmall/Large results (FKGL: lower is better). Note that SARI is the primary, human-correlated metric for sentence simplification (Xu et al., 2016).

Models	HUMAN EVALUATION Fluency	Adequacy	Simplicity	Average	MATCH-WITH-INPUT BLEU (%)	ROUGE (%)	Exact Match (%)
Ground-truth	4.97	4.08	3.83	4.29	18.25	43.74	0.00
Hybrid	3.88	3.82	3.92	3.87	25.74	56.20	3.34
DRESS-LS	4.84	4.18	3.21	4.08	42.93	67.61	14.48
Pointer Baseline	4.61	3.94	3.99	4.18	30.80	60.56	10.68
3-way Multi-task	4.73	3.18	4.62	4.18	8.74	37.82	2.41

Table 4: Human evaluation results (on left) and closeness-to-input source results (on right), for Newsela. In Sec. 5 'Human Evaluation', we discuss the issue of high adequacy scores for outputs that are very similar to the input (see right part of the table).

Models	BLEU	FKGL	SARI
NEWSELA			
Static Mixing Ratio	11.86	1.38	32.98
Dynamic Mixing Ratio	11.14	1.32	33.22
WIKISMALL			
Static Mixing Ratio	29.70	6.93	28.24
Dynamic Mixing Ratio	27.23	5.86	29.58

Table 3: Results on dynamic vs. static mixing ratio (FKGL: lower is better).

Models	Entailment	Paraphrasing
Ground-truth	N/A	62.1
Hybrid	34.8	74.1
DRESS-LS	30.7	77.9
Pointer Baseline	36.9	76.6
2-way Multi-Task	41.4	63.9

Table 6: Analysis: Entailment and paraphrase classification results (avg. probability scores as %) on Newsela.

Models	BLEU	FKGL	SARI
Final (High Ent + Low PP)	**11.86**	**1.38**	**32.98**
Both lower-layer	11.94	1.47	31.92
Both higher-layer	12.26	1.38	32.02
Swapped (Low Ent + High PP)	21.64	2.97	29.07
Hard-sharing	13.01	1.38	32.36

Table 5: Multi-task layer ablation results on Newsela.

Models	Deletions	Additions
Hybrid	95.18	0.000
DRESS-LS	85.37	0.047
Pointer Baseline	88.91	0.026
3-way Multi-Task	97.54	0.049

Table 7: Analysis: SARI's sub-operation scores on Newsela dataset.

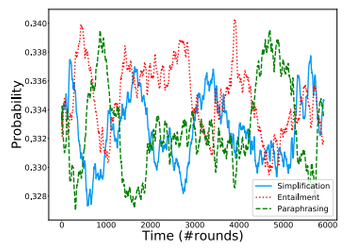

Figure 3: Task selection probability over training trajectory, predicted by bandit controller.

Input: *he put henson in charge of escorting his slaves to his brother 's kentucky plantation .*
Reference: *he sent henson to take his slaves to kentucky .*
DRESS-LS: *he put henson in charge of escorting his slaves to his brother 's kentucky plantation .*
Baseline: *he put his slaves to his brother 's kentucky plantation .*
Multi-Task: *he put henson in charge of escorting .*

Input: *northern states did not allow slavery , but escaped slaves were returned to their owners as property , so henson would have to flee to canada to be free .*
Reference: *states in the north did not allow slavery .*
DRESS-LS: *southern states did not allow slavery , but the guatemalans were returned to their owners as property .*
Baseline: *he slaves were returned to their owners as property .*
Multi-Task: *northern states did not allow slavery .*

Figure 4: Output examples comparing DRESS-LS, our pointer baseline, and multi-task model.

Acknowledgments:
We thank the reviewers for their helpful comments (and Xingxing Zhang for providing preprocessed datasets). This work was supported by DARPA (YFA17-D17AP00022), Google Faculty Research Award, Bloomberg Data Science Research Grant, and Nvidia GPU awards. The views contained in this article are those of the authors and not of the funding agency.

References:

- See, Liu, and Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.
- Bowman, Angeli, Potts, and Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.
- Williams, Nangia , and Bowman . 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*.
- Wieting, Gimpel. 2018. ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. In *ACL*.
- Bahdanau, Cho, and Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.
- Belinkov, Durrani, Dalvi, Sajjad, and Glass. 2017. What do neural machine translation models learn about morphology? In *ACL*.
- Xu, Napoles, Pavlick, Chen, and Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification In *TACL*.
- Pasunuru and Bansal. 2017. Multi-task video captioning with video and entailment generation. In *ACL*.
- Graves, Bellemare, Menick, Munos, and Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *ICML*.
- Sharma and Ravindran. 2017. Online multi-task learning using active sampling. In *ICLR Workshop*.