Investigating Softmax Tempering for Training Neural Machine Translation Models

Raj Dabre and Atsushi Fujita
NICT, Japan

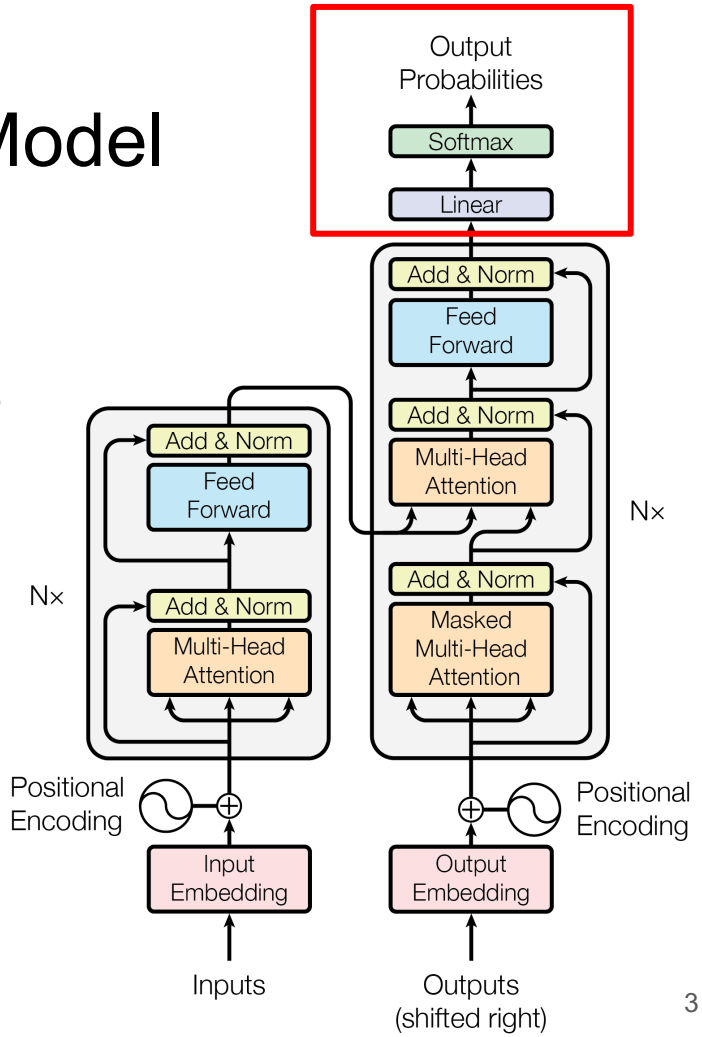
MT Summit
16-20 August, 2021

Overview

- Current solutions for NMT overfitting
  - Leveraging data: multilingual training, back-translation, fine-tuning pre-trained models
  - Parameter regularization: weight regularization, dropout, etc.
  - Hyper-parameter tuning
- Key observation: few softmax centric regularization methods
  - e.g., label smoothing, entropy maximization
  - This is where the observable action occurs
- Our solution: softmax tempered training
  - Smooth softmax using temperature to make it harder to overfit
  - **Result: Improved BLEU especially for low-resource conditions**

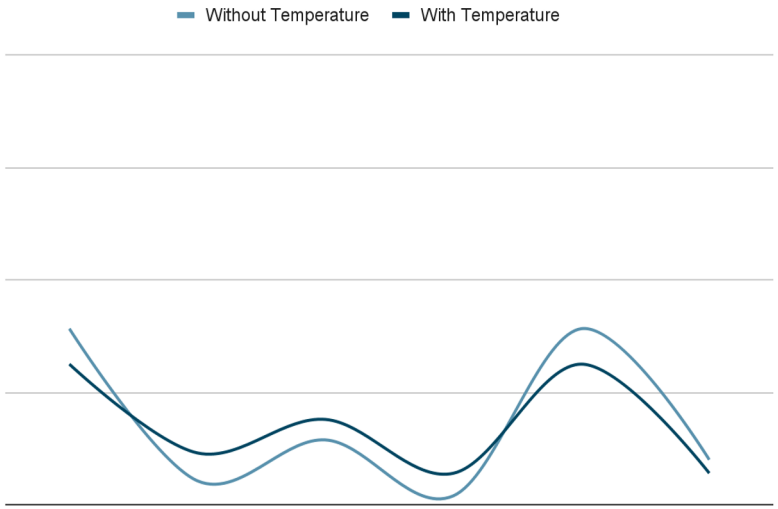
Background: Transformer NMT Model

- Attention based encoder-decoder
  - Stacked encoder and decoder
- Objective: minimize softmax cross entropy loss
  - Probability of predicting *i*-th word:
    - $P_i = P(Y_i \mid Y_{<i}, X) = \text{softmax}(D_i)$
    - $D_i \in \mathbf{R}^V$: the logit for ith word
    - $P_i \in [0, 1]^V$
  - Loss:
    - $L_i = -\langle \log(P_i), R_i \rangle$
    - $\langle .,. \rangle$ is the inner product
    - $R_i \in [0, 1]^V$: the (smoothed) label for *i*-th word
- Issue:
  - No explicit regularization on softmax



Our Approach: Softmax Tempering

- Softmax tempered prediction:
  - $P_{temp} = \text{softmax}(D_i/T)$, where $T > 1$
  - Multiply loss by T for gradient rescaling
  - Used for knowledge distillation (Hinton+, 2015)
- Effects:
  - Smoother softmax
  - Model will be prevented from easily overfitting
- Flow of training:
  - Temperature makes smoothed softmax
  - Loss forces model to produce sharper softmax next time
  - Temperature thwarts this attempt
  - Endless cycle till convergence
- Beam decoding should be done with temperature



Experiments: Datasets

- Low-resource settings: Asian Language Treebank (ALT)
  - 18,088 / 1,000 / 1,018 training / development / testing pairs
  - 11 Asian languages: Bengali (Bn), Filipino (Fil), Indonesian (Id), Japanese (Ja), Khmer (Km), Lao (Lo), Malay (Ms), Burmese (My), Thai (Th), Vietnamese (Vi), and Chinese (Zh)
  - English-to-Asian and Asian-to-English translation
- High-resource setting: WMT 2019 English to German translation
  - See paper for more details
- Preprocessing:
  - Segmentation of Ja, Zh, Km, Lo, My, and Th using an in-house segmenter

Experiments: Implementation

- tensor2tensor v1.14
  - Transformer Base for ALT dataset (vocab: 8,192)
  - **IMPORTANT: Label smoothing of 0.1 by default**
  - Temperatures tested: 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 3.0, 4.0, 5.0, and 10.0.
  - Training till convergence
- Evaluation criteria: BLEU (Papineni+, 2002)
- Choosing model for decoding:
  - Average last 10 checkpoints
  - Decode with the temperature used during training
  - Optimal temperature: model with best greedy development score
    - Greedy and beam scores are correlated
  - Optimal beam size and length penalty: grid search using optimal temperature model
    - Beam sizes (7): 1, 2, 4, 6, 8, 10, and 12
    - Length penalties (9): 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, and 1.4

Results: Low-Resource Settings

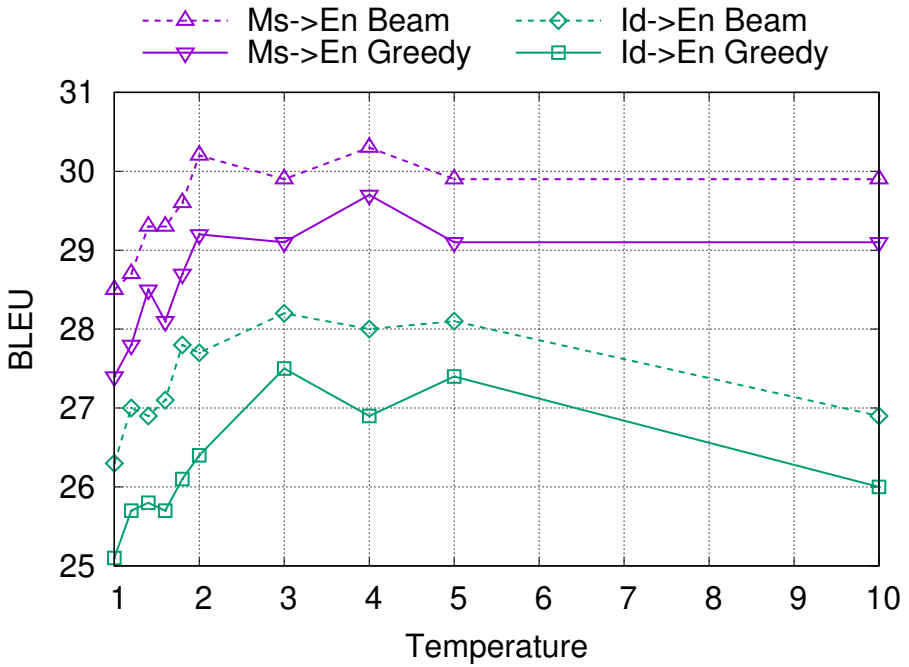
T	Decoding	En→XX Bn	Fil	Id	Ja	Km	Lo	Ms	My	Th	Vi	Zh
1.0	Greedy	3.5	24.3	27.4	13.4	19.3	11.5	31.5	8.3	13.7	24.0	10.4
1.0	Beam	4.1	25.8	28.7	15.0	21.3	13.0	32.6	9.1	15.9	26.5	12.1
T_{opt}	Greedy	4.5	25.7	29.5†	15.5	20.7	11.8	33.7†	9.3	15.6	25.8	12.9†
T_{opt}	Beam	**4.7**	**27.0**†	**30.2**†	**17.5**†	**22.3**†	**13.3**†	**34.7**†	**10.6**†	**17.4**†	**27.5**†	**15.1**†
Value for T_{opt}		5.0	3.0	4.0	4.0	5.0	5.0	4.0	5.0	5.0	3.0	5.0

T	Decoding	XX→En Bn	Fil	Id	Ja	Km	Lo	Ms	My	Th	Vi	Zh
1.0	Greedy	7.1	22.2	25.1	8.7	14.9	9.8	27.4	7.8	10.5	19.4	9.4
1.0	Beam	8.5	24.0	26.3	9.9	16.4	11.9	28.5	9.3	12.4	20.9	10.8
T_{opt}	Greedy	9.1	24.7	27.5†	11.0†	16.8	11.4	29.7†	11.7†	12.2	21.3	11.5
T_{opt}	Beam	**10.4**†	**26.3**†	**28.2**†	**12.9**†	**18.0**†	**12.9**†	**30.3**†	**13.3**†	**13.7**†	**22.1**†	**12.9**†
Value for T_{opt}		5.0	5.0	3.0	5.0	4.0	5.0	4.0	4.0	4.0	4.0	5.0

- Softmax tempering improves performance (duh!)
  - Temperature values around 3.0 to 5.0 work the best

Result: BLEU Variation with Temperature

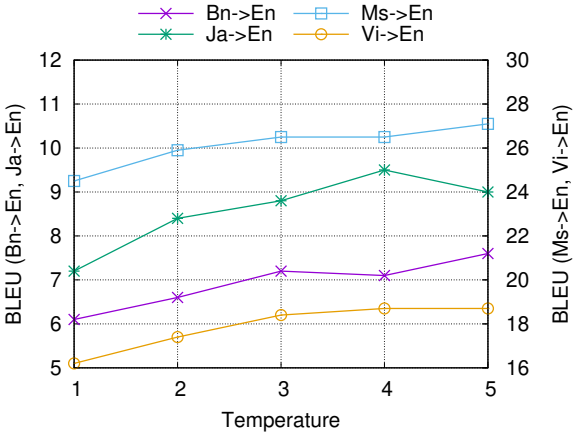
- Test set greedy-beam scores
  - Malay/Indonesian-to-English
  - Vary temperatures from 1.0 to 10.0
- $1.0 < T <= 2.0$
  - Small positive improvement
- $2.0 < T <= 5.0$
  - Large improvements
  - Peak around 3.0 to 5.0
- $T > 5.0$
  - Drop in performance
  - Overtly smoothed softmaxes
  - Overtly large gradients

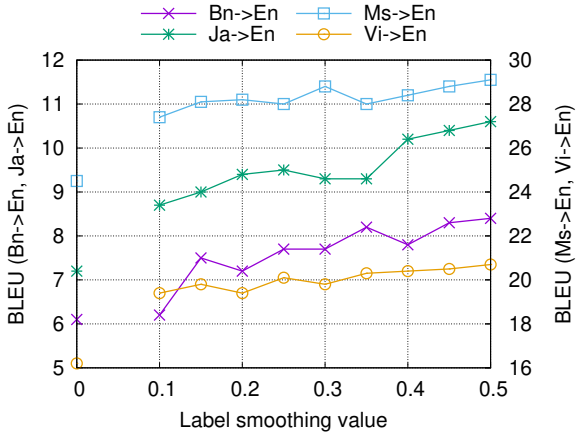


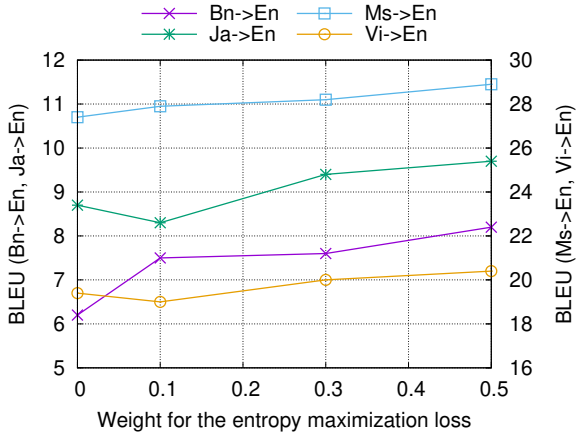
Relationship with Other Smoothing Approaches

- Softmax Tempering (Ours)
  - Original: $P = \mathrm{softmax}(D_i)$
  - Tempered: $P_{temp} = \mathrm{softmax}(D_i/T)$
  - Smooth softmax directly
- Label Smoothing (Szegedy+, 2016)
  - Original: R_i = {1 if V_i is correct label else 0} (One hot label)
  - Smoothed: R_i = {(1-S)+S/V if V_i is correct label else S/V} (S: the smoothing factor)
  - Smoothed label forces smoother softmax
- Softmax Entropy Maximization (Pereyra+, 2017)
  - Softmax cross entropy: $-\langle \log(P_i), R_i \rangle$
  - Softmax entropy: $-\langle \log(P_i), P_i \rangle$
  - Their linear combination: $-(1-w)*\langle \log(P_i), R_i \rangle$ - w*$\langle \log(P_i), P_i \rangle$ (w: the weight)
- Main question: are these complementary?

Each method improves BLEU





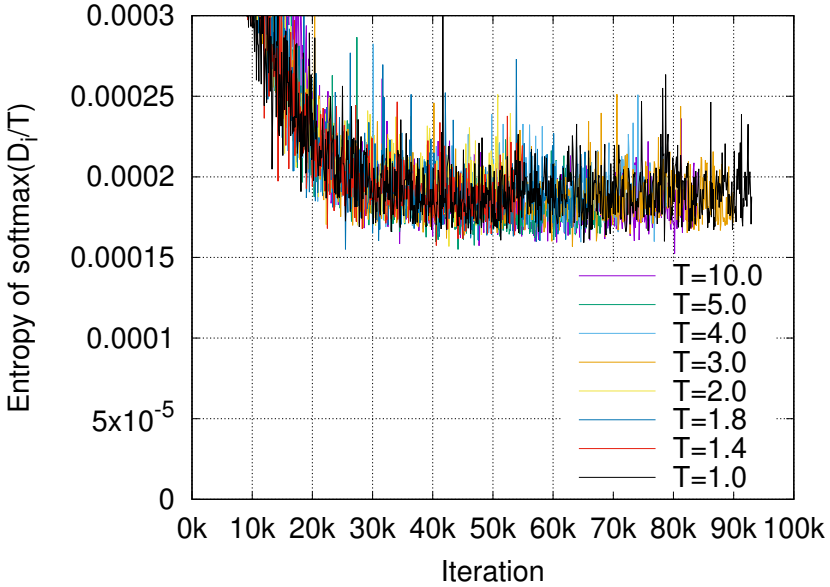


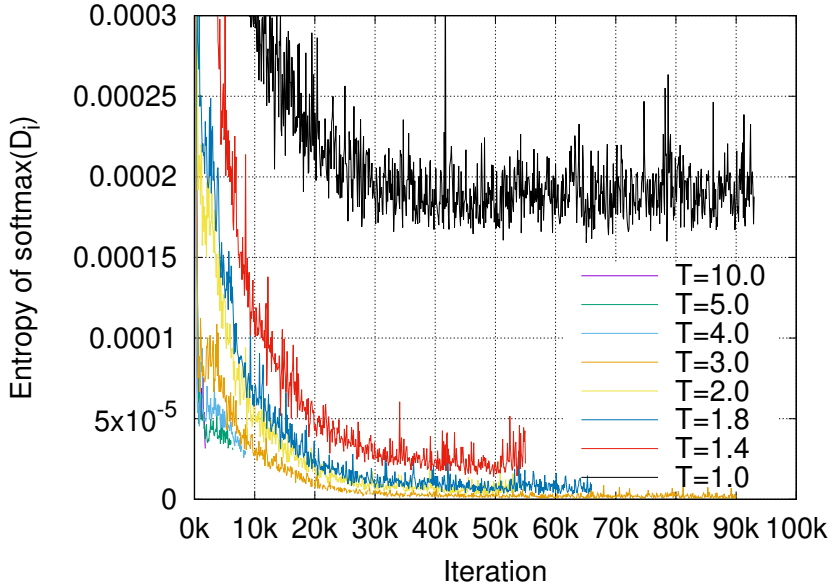
How about all together?

Combined methods			Bn→En		Ja→En		Ms→En		Vi→En	
Temp	LS	SEM	BLEU	(*T*, *S*, *w*)	BLEU	(*T*, *S*, *w*)	BLEU	(*T*, *S*, *w*)	BLEU	(*T*, *S*, *w*)
-	-	-	7.1	(1.0, 0.1, 0.0)	8.7	(1.0, 0.1, 0.0)	27.4	(1.0, 0.1, 0.0)	19.4	(1.0, 0.1, 0.0)
-	✓	-	8.4	(1.0, 0.5, 0.0)	10.6	(1.0, 0.5, 0.0)	29.1	(1.0, 0.5, 0.0)	20.7	(1.0, 0.5, 0.0)
✓	-	-	9.1	(5.0, 0.1, 0.0)	11.0	(5.0, 0.1, 0.0)	29.7	(4.0, 0.1, 0.0)	21.3	(4.0, 0.1, 0.0)
✓	✓	-	**9.4**	(5.0, 0.5, 0.0)	**11.8**	(5.0, 0.5, 0.0)	**30.1**	(5.0, 0.45, 0.0)	**22.1**	(4.0, 0.45, 0.0)
-	-	✓	8.2	(1.0, 0.1, 0.5)	9.7	(1.0, 0.1, 0.5)	28.9	(1.0, 0.1, 0.5)	20.4	(1.0, 0.1, 0.5)
-	✓	✓	8.8	(1.0, 0.5, 0.1)	11.2	(1.0, 0.5, 0.3)	28.9	(1.0, 0.5, 0.3)	20.8	(1.0, 0.3, 0.3)
✓	-	✓	8.8	(5.0, 0.1, 0.5)	11.5	(5.0, 0.1, 0.5)	**30.1**	(5.0, 0.1, 0.5)	21.6	(5.0, 0.1, 0.1)
✓	✓	✓	**9.4**	(5.0, 0.5, 0.0)	**11.8**	(5.0, 0.5, 0.0)	**30.1**	(5.0, 0.45, 0.0)	**22.1**	(4.0, 0.45, 0.0)

- Not completely orthogonal
  - Tempering and label smoothing invalidate entropy maximization
- BUT: Do experiment with all the combinations

Whats Under The Hood?





- Left: Softmax entropy during training with temperature
- Right: Softmax entropy during training without temperature
- Tempered training leads to sharper untempered distributions (low-entropy)

Conclusion

- Softmax tempering
  - Improved BLEU on low-resource settings
  - Limited impact on high-resource languages
  - Visualization reveals eventual sharpening of softmax
- A study of softmax tempering, label smoothing, and entropy maximization
  - All approaches are pairwise compatible
  - Entropy maximization is least helpful
- Future work
  - Automatic temperature learning
  - Optimal temperature for various corpora sizes