Paper

Latent Alignment and Variational Attention

Yuntian Deng* **Yoon Kim*** **Justin Chiu** **Demi Guo** **Alexander M. Rush**

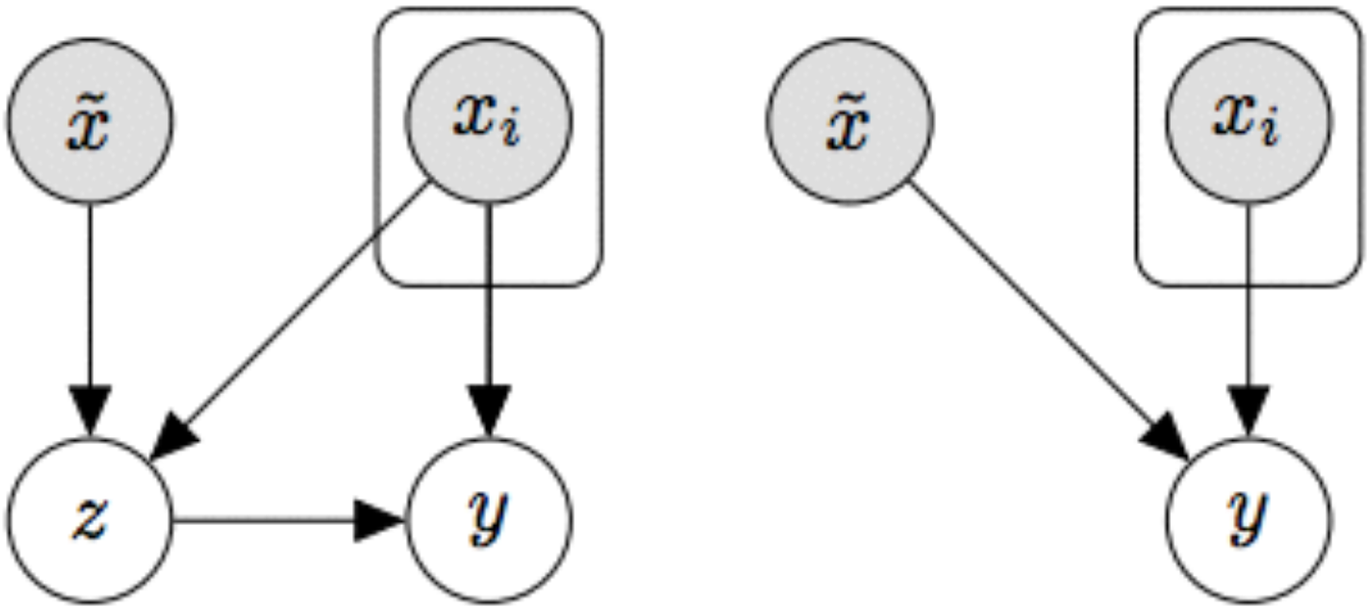
{dengyuntian@seas,yoonkim@seas,justinchiu@g,dguo@college,srush@seas}.harvard.edu

School of Engineering and Applied Sciences
Harvard University
Cambridge, MA, USA

Diff between Soft/Hard Attention

- Soft Attention
  - Encoding inductive biases (necessary assumption)
  - interpretability
- Hard Attention
  - Introducing an alignment explicitly
  - Optimizing a bound on log marginal llh using PG.
  - Performing worse than soft attention.

Background



(a) (b)

Figure 2: Models over observed set x, query $\tilde{x}$, and alignment z. (a) Latent alignment model, (b) Soft attention with z absorbed into prediction network.

Advantages of Latent Alignment

- Facilitate reasoning about dependencies in a prob way.
- Posterior inference helps model analysis than FF models.
- Directly maximising marginal llh providing better results.

Latent Alignment Models

- Two steps generation:

$$z \sim \mathcal{D}(a(x, \tilde{x}; \theta)) \quad y \sim f(x, z; \theta)$$

- Learning

$$\max_{\theta} \log p(y = \hat{y} \,|\, x,\ \tilde{x}) \quad = \quad \max_{\theta} \log \mathbb{E}_z[f(x, z; \theta)_{\hat{y}}]$$

$$f(x, z) = \text{softmax}(\mathbf{W} X z)$$

- Two distributions for z:
  - Categorical distribution: z is a one hot vector

$$\log p(y = \hat{y} \,|\, x,\ \tilde{x}) = \log \sum_{i=1}^{T} p(z_i = 1 \,|\, x, \tilde{x}) p(y = \hat{y} \,|\, x, z_i = 1) = \log \mathbb{E}_z[\text{softmax}(\mathbf{W} X z)_{\hat{y}}]$$

  - Relaxed alignments: intractable to compute exactly.

Soft Attentions

- Soft Attention

$$\log p_{\text{soft}}(y \mid x,\ \tilde{x}) = \log f(x, \mathbb{E}_z[z]; \theta) = \log \text{softmax}(\mathbf{W} X \mathbb{E}_z[z])$$

- Bound on Soft Attention

Proposition 1. *Define* $g_{x,\hat{y}} : \Delta^{T-1} \mapsto [0,1]$ *to be the function given by* $g_{x,\hat{y}}(z) = f(x,z)_{\hat{y}}$ *(i.e.* $g_{x,\hat{y}}(z) = p(y = \hat{y} \mid x, \tilde{x}, z)$*) for a twice differentiable function* f*. Let* $H_{g_{x,\hat{y}}}(z)$ *be the Hessian of* $g_{x,\hat{y}}(z)$ *evaluated at* z*, and further suppose* $\|H_{g_{x,\hat{y}}}(z)\|_2 \leq c$ *for all* $z \in \Delta^{T-1}, \hat{y} \in \mathcal{Y}\}$*, and* x*, where* $\|\cdot\|_2$ *is the spectral norm. Then for all* $\hat{y} \in \mathcal{Y}$*,*

$$|p(y = \hat{y} \mid x, \tilde{x}) - p_{\text{soft}}(y = \hat{y} \mid x, \tilde{x})| \leq c$$

- Empirically soft attention works very well in practice.
- Soft attention moves towards a sharper distribution
- Alignments learned are corrected with human intuition.

Hard Attentions

- Hard Attention
- Works by two steps:
  - Computing lower bound on the log marginal llh
  - Maximising low bound by policy gradient

$$\nabla_\theta \mathbb{E}_z[\log f(x,z))] = \mathbb{E}_z[\nabla_\theta \log f(x,z) + (\log f(x,z) - B)\nabla_\theta \log p(z \mid x, \tilde{x})]$$

Variational Attention

- Motivated from hard attention that the gap derived from the jensen's inequality is large, contributing to poor acc.
- Variational inference methods aim to tighten this gap.

$$\log \mathbb{E}_{z\sim p(z\,|\,x,\tilde{x})}[p(y\,|\,x,z)] \geq \mathbb{E}_{z\sim q(z)}[\log p(y\,|\,x,z)] - \mathrm{KL}[q(z)\,\|\,p(z\,|\,x,\tilde{x})]$$

- The bound is tight when q(z)=p(z|x,x',y).
- Hard attention is a special case of ELBO: q(z)=p(z|x,x').
- q(z) is modelled by an inference network.

$$q(z;\lambda) \qquad\qquad \lambda = enc(x,\tilde{x},y;\phi)$$

Variational Attention

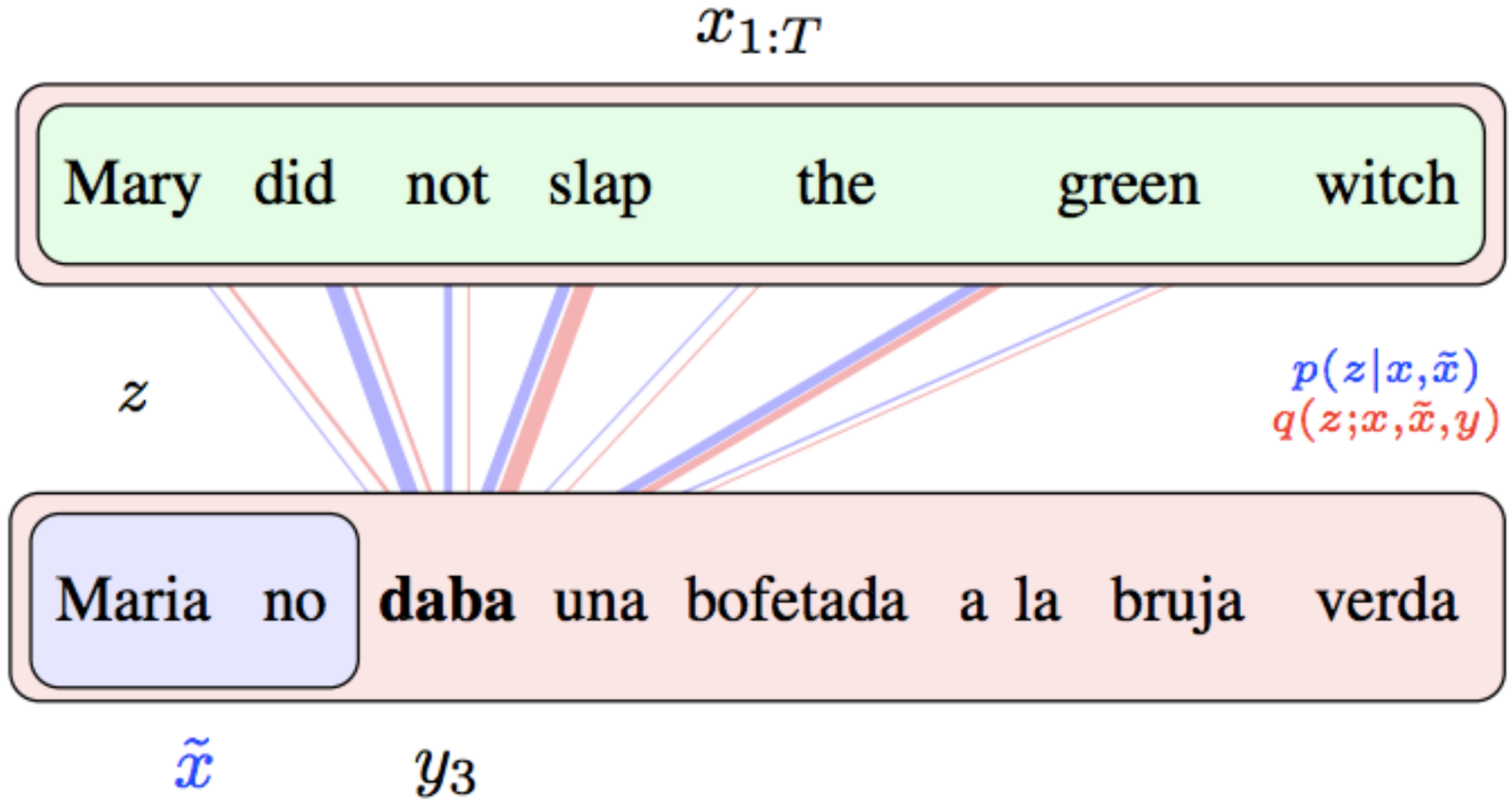


Figure 1: Sketch of variational attention applied to neural machine translation. Two alignment distribution are shown, the blue prior p computed using known information, and the red variational posterior q taking into account future observations. Our aim is to use q to improve estimates of p and to support improved inference of z.

Algo for Categorical Alignments

- Gradient for the KL term is easy.
- For the first term: apply policy gradient.

Algo for Relaxed Alignments

- D and Q are Dirichlets distributions.
- Instead of using REINFORCE algorithm, gradient is computed by reparameterization trick.

$$\mathbb{E}_{u\sim\mathcal{U}}\left[\nabla_\phi \log p(y|x, g_\phi(u))\right] - \nabla_\phi \,\mathrm{KL}\left[q(z) \,\|\, p(z \mid x, \tilde{x})\right]$$

Predictive Inference

- For categorical distribution:
  - Direct enumerating
  - K-max approximation
- For relaxed case: sampling is necessary.

Experiments

- Two tasks: NMT and Visual QA
- Experiments vary three components:
  - training objective and model
  - training approximations
  - test inference
- Besides this, all the neural models have the same architecture and exact same parameters.

Experiment 1

Model	Objective	$\mathbb{E}$	NMT PPL	NMT BLEU	VQA NLL	VQA Eval
Soft Attention	$\log p(y \mid \mathbb{E}[z])$	-	7.03	32.31	1.76	58.93
Marginal Likelihood	$\log \mathbb{E}[p]$	Enum	6.33	33.08	1.69	60.33
Hard Attention	$\mathbb{E}_p[\log p]$	Enum	7.37	31.40	1.78	57.60
Hard Attention	$\mathbb{E}_p[\log p]$	Sample	7.38	31.00	1.82	56.30
Variational Relaxed Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	7.58	30.05	-	-
Variational Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Enum	6.03	33.10	1.69	58.44
Variational Attention	$\mathbb{E}_q[\log p] - \mathrm{KL}$	Sample	6.13	33.09	1.75	57.52

Table 1: Evaluation on neural machine translation (NMT) and visual question answering (VQA) for the various models. $\mathbb{E}$ column indicates whether the expectation is calculated via enumeration (Enum) or a single sample (Sample) during training. For NMT we evaluate intrinsically on perplexity (PPL) (lower is better) and extrinsically on BLEU (higher is better), where for BLEU we perform beam search with beam size 10 and length penalty (see Appendix B for further details). For VQA we evaluate intrinsically on negative log-likelihood (NLL) (lower is better) and extrinsically on VQA evaluation metric (higher is better). All results except for relaxed attention use enumeration at test time.

- Hard attention performs worse than soft attention.
- Exact marginal llh outperforms soft attention.
- Variational attention performs comparably to optimizing explicit log marginal llh.

Experiments 2

Model	PPL Exact	PPL K-Max	BLEU Exact	BLEU K-Max
Marginal Likelihood	6.33	6.89	33.08	32.97
Hard + Enum	7.37	7.37	31.40	31.37
Hard + Sample	7.38	7.38	31.00	31.04
Variational + Enum	6.03	6.37	33.10	33.00
Variational + Sample	6.13	6.45	33.09	33.01

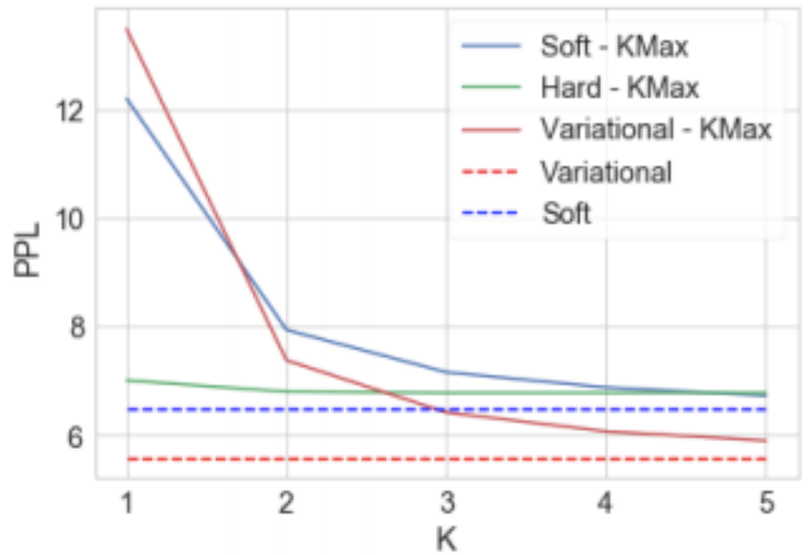
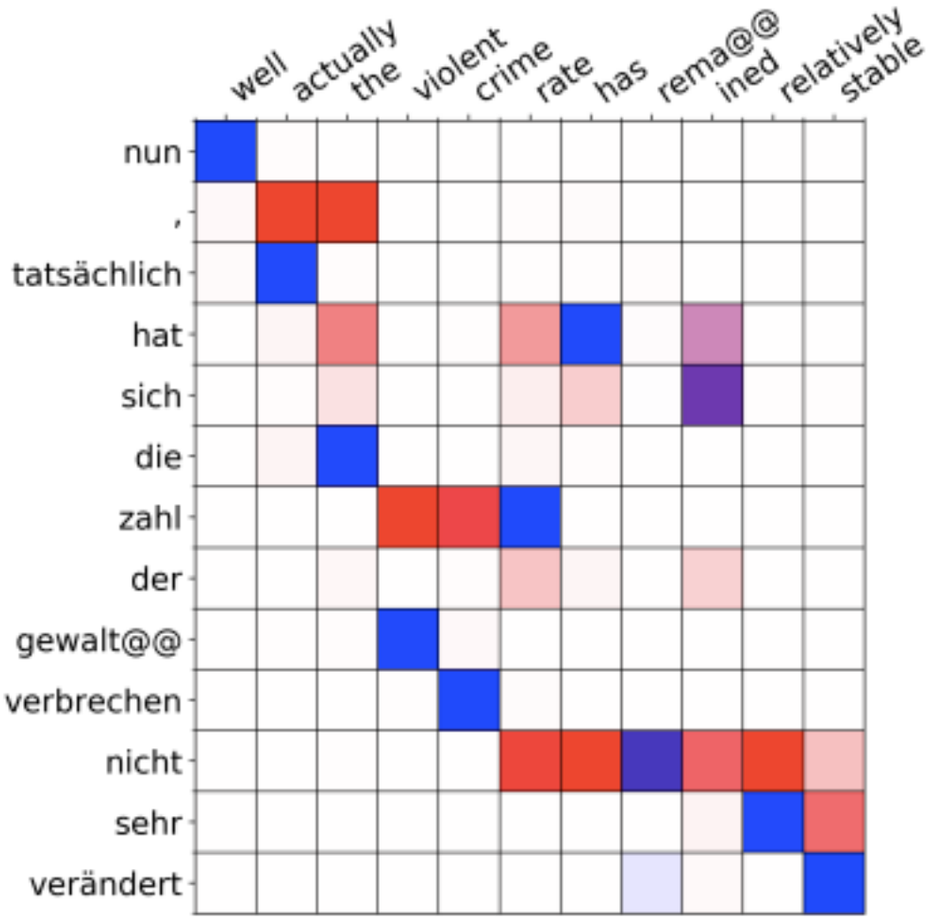


Table 2: (Left) Performance change on NMT from exact decoding to K-Max decoding with $K = 5$. (see section 5 for definition of K-max decoding). (Right) Test perplexity of different approaches while varying the number of k-max samples to estimate $\mathbb{E}_z[p(y|x, \tilde{x})]$. Dotted lines compare soft baseline and variational with full enumeration.

- Exact enumeration is better.
- K-max approximation is reasonable.
- It is possible to train with soft attention and test using K-Max with a small performance drop.
- This possibly indicates that soft attention models are indeed approximating latent alignment models.

Experiments 3



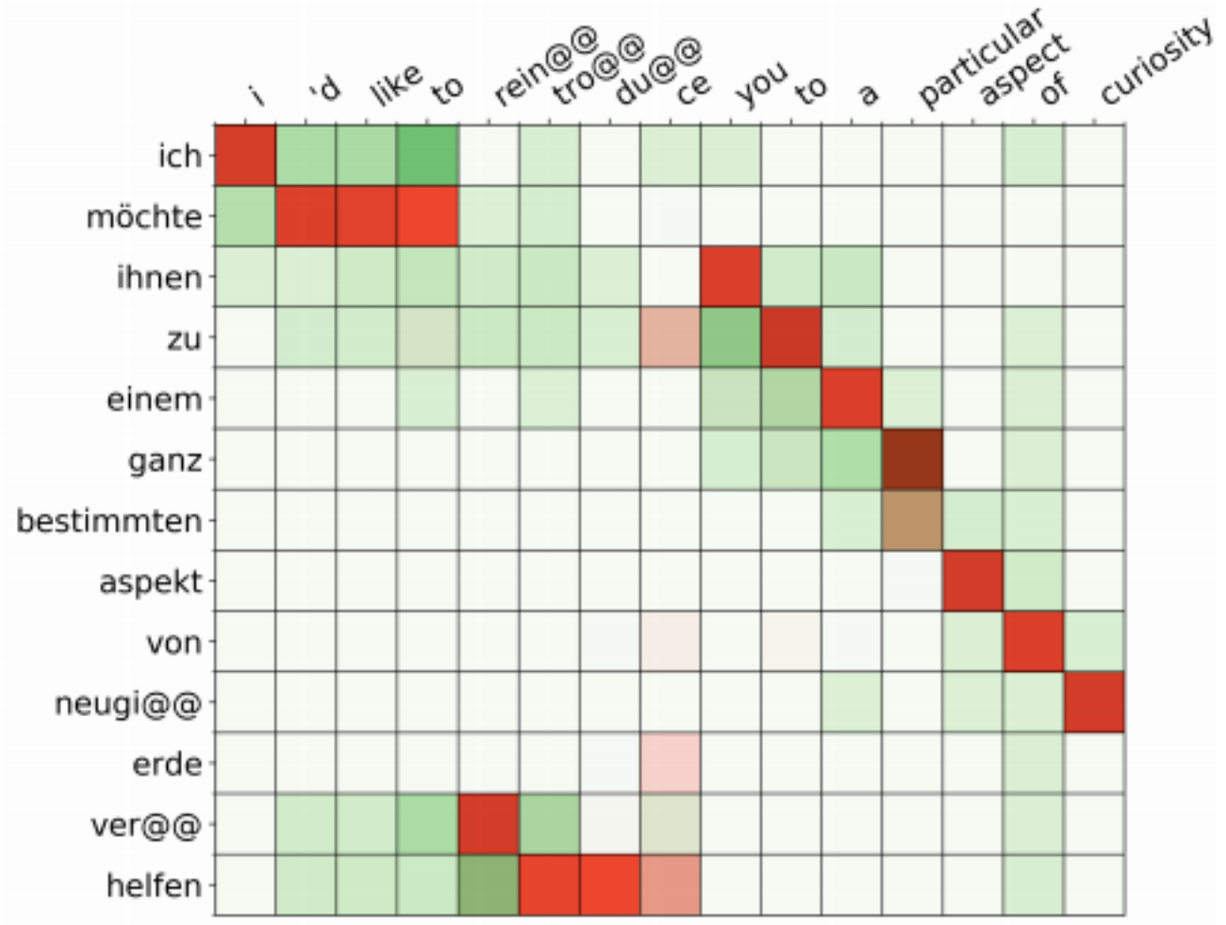


Figure 3: (Left) An example demonstrating the difference between the prior alignment (red) and the variational posterior (blue) when translating from DE-EN (left-to-right). Note the improved blue alignments for `actually` and `violent` which benefit from seeing the next word. (Right) Comparison of soft attention (green) with the p of variational attention (red). Both models imply a similar alignment, but variational attention is lower entropy.

- Left: foresight improves attention.
- Right: entropy of variational attention is low.

Experiments 4

Model	BLEU
Beam Search Optimization [67]	26.36
Actor-Critic [5]	28.53
Neural PBMT + LM [26]	30.08
Minimum Risk Training [19]	32.84
Soft Attention	32.31
Marginal Likelihood	33.08
Hard Attention + Enum	31.40
Hard Attention + Sample	30.42
Variational Relaxed Attention	30.05
Variational Attention + Enum	33.10
Variational Attention + Sample	33.09

Model	Entropy NMT	Entropy VQA
Soft Attention	1.24	2.70
Marginal Likelihood	0.82	2.66
Hard Attention + Enum	0.05	0.73
Hard Attention + Sample	0.07	0.58
Variational Relaxed Attention	2.02	-
Variational Attention + Enum	0.54	2.07
Variational Attention + Sample	0.52	2.44

Table 3: (Left) Comparison against the best prior work for NMT on the IWSLT 2014 German-English test set. (Right) Comparison of different models in terms of implied discrete entropy (lower = more peaked alignment).

- Variational relaxed attention performs worst.
- Extreme low entropy of variational relaxed attention

Paper

AUTO-ENCODING VARIATIONAL NEURAL MACHINE TRANSLATION

Bryan Eikema & Wilker Aziz
Institute for Logic, Language and Computation
University of Amsterdam
bryan.eikema@student.uva.nl, w.aziz@uva.nl

Motivation

- Training data of bilingual corpus is mixed.
  - Multiple sources
  - Different domains
- Current NMT models are conditional models.

Neural Machine Translation

- Training: likelihood function.

$$P(y|x,\theta) = \prod_{j=1}^{|y|} P(y_j|x, y_{<j}, \theta)$$

$$= \prod_{j=1}^{|y|} \mathrm{Cat}(y_j|f_\theta(x, y_{<j}))$$

- Prediction:

$$\arg\max_y \ \log P(y|x,\theta) \approx \mathrm{greedy}_y \ \log P(y|x,\theta)$$

Auto-encoding Variational NMT

- Generative story,

$$p(z, x, y|\theta) = p(z)P(x, y|z, \theta)$$

- Second term:

$$P(x, y|z, \theta) = \prod_{i=1}^{|x|} P(x_i|z, x_{<i}, \theta) \prod_{j=1}^{|y|} P(y_j|z, x, y_{<j}, \theta)$$

$$= \prod_{i=1}^{|x|} \mathrm{Cat}(x_i|g_\theta(z, x_{<i})) \prod_{j=1}^{|y|} \mathrm{Cat}(y_j|f_\theta(z, x, y_{<j}))$$

Auto-encoding Variational NMT

- Statistical consideration,
  - The distribution over source sentences can provide no information about the conditional distribution over target sentences given a source.

$$P(x, y|z, \theta) = P(x|\underbrace{z, \theta_{\text{emb-x}}, \theta_{\text{LM}}}_{\alpha})P(y|x, \underbrace{z, \theta_{\text{emb-x}}, \theta_{\text{TM}}}_{\beta})$$

- Learning:
  - ELBO

$$\log P(x, y|\theta) \geq \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\log P(x, y|Z = \mathbf{u} + \epsilon \odot \mathbf{s}, \theta)\right] - \mathrm{KL}(\mathcal{N}(z|\mathbf{u}, \mathrm{diag}(\mathbf{s} \odot \mathbf{s}))||\mathcal{N}(z|0, I)) \ ,$$

Auto-encoding Variational NMT

- Prediction:

$$
\begin{aligned}
\arg\max_y \log P(y|x) &= \arg\max_y \log P(y,x) \\
&\approx \arg\max_y \mathbb{E}_{q(z|x,y)}[\log P(y,x|Z)] - \mathrm{KL}(q(z|x,y)||p(z)) \\
&\approx \arg\max_y \mathbb{E}_{r(z|x)}[\log P(y|Z,x) + \log P(x|Z)] - \mathrm{KL}(r(z|x)||p(z)) \\
&= \arg\max_y \mathbb{E}_{r(z|x)}[\log P(y|Z,x)] \\
&\approx \arg\max_y \log P(y|\mathbb{E}_{r(z|x)}[Z],x)
\end{aligned}
$$

- Auxiliary distribution:
  - Choice 1: q(z|x,y)=q(z|x)
  - Choice 2: modified ELBO

$$\log P(x,y|\theta) \geq \mathrm{ELBO}(\theta,\lambda|x,y) - \mathrm{D}(r_\phi, q_\lambda)$$

Experiments

- Two translation tasks:
  - WMT's translation of news
  - IWSLT's translation of transcripts of TED talks
- German (DE) and English (EN) in either direction
- Three experiments:
  - In-domain
  - Mixed-domain training
  - Corpus with synthetic data

Results on In-domain

IWSLT14	EN-DE		DE-EN	
	BLEU ↑	BEER ↑	BLEU ↑	BEER ↑
CONDITIONAL	23.4	59.1	28.7	60.6
JOINT	23.6	59.2	28.7	60.6
AEVNMT	**23.9**	**59.4**	**29.3**	**61.0**

Table 3: Test results for in-domain training on IWSLT.

WMT16	EN-DE		DE-EN	
	BLEU ↑	BEER ↑	BLEU ↑	BEER ↑
CONDITIONAL	17.7	53.5	20.5	**54.3**
JOINT	18.0	**53.8**	20.5	54.2
AEVNMT	**18.4**	53.7	**20.8**	54.1

Table 4: Test results for in-domain training on NC.

Results on Mix-domain

EN-DE		WMT16		IWSLT14	
Training	Model	BLEU ↑	BEER ↑	BLEU ↑	BEER ↑
In-Domain	CONDITIONAL	17.7	53.5	23.4	59.1
	AEVNMT	18.4	53.7	23.9	59.4
Mixed-Domain	CONDITIONAL	17.3	54.2	24.1	59.7
	AEVNMT	**18.6**	**55.1**	**24.2**	**59.8**

Table 5: EN-DE test results for mixed-domain training. In-domain results copied from Tables 3 and 4 for comparison.

DE-EN		WMT16		IWSLT14	
Training	Model	BLEU ↑	BEER ↑	BLEU ↑	BEER ↑
In-Domain	CONDITIONAL	20.5	54.3	28.7	60.6
	AEVNMT	20.8	54.1	29.3	61.0
Mixed-Domain	CONDITIONAL	22.2	55.7	**30.6**	61.8
	AEVNMT	**22.7**	**56.2**	30.5	**61.9**

Table 6: DE-EN test results for mixed-domain training. In-domain results copied from Tables 3 and 4 for comparison.

Results on synthetic-domain

WMT16	EN-DE		DE-EN	
	BLEU ↑	BEER ↑	BLEU ↑	BEER ↑
CONDITIONAL	17.7	53.5	20.5	54.3
+ synthetic data	22.3	57.4	27.0	59.0
JOINT + synthetic data	22.3	57.5	27.2	59.1
AEVNMT + synthetic data	**22.6**	**57.6**	**27.9**	**59.3**

Table 7: Test results for training on NC plus synthetic data (back-translated News Crawl).