Modelling Latent Variation in Translation Data

Wilker Aziz
University of Amsterdam
(joint work with Philip Schulz and Trevor Cohn)

May 25, 2018

Outline

1. Variation in translation data
2. Neural machine translation
3. Deep generative MT
4. Experiments
5. Remarks

Variation in translation data

There are latent factors of variation in translation data

Variation in translation data

There are latent factors of variation in translation data

Eles **realizaram um estudo** *sobre a prevalência de autismo na população* ...

- They performed a study *on* the prevalence of autism in the general population ...
- They undertook a study *of* autism prevalence in the general population ...
- They studied *the widespread of* autism in the general population ...

Variation in translation data

There are latent factors of variation in translation data

Eles **realizaram um estudo** *sobre a prevalência de autismo na população* ...

- They performed a study *on* the prevalence of autism in the general population ...
- They undertook a study *of* autism prevalence in the general population ...
- They studied *the widespread of* autism in the general population ...
- In a study on the prevalence of autism in the population conducted by FAPESP ...

Where does variation come from?

Bhagat and Hovy (2013) identified more than 20 linguistic devices!

Where does variation come from?

Bhagat and Hovy (2013) identified more than 20 linguistic devices!

Some of the "simple" ones:

- Synonym substitution
  Google bought YouTube $\Leftrightarrow$ Google acquired YouTube

Where does variation come from?

Bhagat and Hovy (2013) identified more than 20 linguistic devices!

Some of the “simple” ones:

- Synonym substitution
  Google bought YouTube $\Leftrightarrow$ Google acquired YouTube
- Converse substitution
  Google bought YouTube $\Leftrightarrow$ YouTube was sold to Google

Where does variation come from?

Bhagat and Hovy (2013) identified more than 20 linguistic devices!

Some of the "simple" ones:

- Synonym substitution
  Google bought YouTube ⇔ Google acquired YouTube
- Converse substitution
  Google bought YouTube ⇔ YouTube was sold to Google
- Change of voice
  Google bought YouTube ⇔ YouTube was bought by Google

Can we model it?

Some variation we can hope to model

- synonym/antonym/converse substitution
- change of voice/person
- function word variation
- semantic role substitution
- POS conversion

Bhagat and Hovy (2013): 85% MTC and 60% MSR

Can we model it?

Some of it is far beyond the scope of a sentence pair

- coherence devices
- co-referent substitution

Bhagat and Hovy (2013): 15% MTC and 40% MSR

Can we model it?

Some of it is far beyond the scope of a sentence pair

- coherence devices
- co-referent substitution

Some of it is far beyond the corpus

- external knowledge
  The government/Bush declared victory in Iraq

Bhagat and Hovy (2013): 15% MTC and 40% MSR

Can we model it?

Some of it is far beyond the scope of a sentence pair

- coherence devices
- co-referent substitution

Some of it is far beyond the corpus

- external knowledge
  The government/Bush declared victory in Iraq
- evaluation, connotation, viewpoint
  The school said that their buses seat/cram in 40 students each

Bhagat and Hovy (2013): 15% MTC and 40% MSR

Can we model it?

Some of it is specific to how the data was created

- level of proficiency
- cultural background
- fatigue

Summary

There is latent variation in translation data

Summary

There is latent variation in translation data

- some of which is "rather simple"
  undertook/performed a study ...

Summary

There is latent variation in translation data

- some of which is "rather simple"
  undertook/performed a study ...
- some of which requires planning
  a study on the prevalence of ... conducted by ...

Summary

There is latent variation in translation data

- some of which is "rather simple"
  undertook/performed a study ...
- some of which requires planning
  a study on the prevalence of ... conducted by ...
- some of which can only be thought of as random effects unless we
  - expand beyond sentence-level processing
  - incorporate world knowledge
  - model pragmatics

Summary

There is latent variation in translation data

- some of which is "rather simple"
  undertook/performed a study ...
- some of which requires planning
  a study on the prevalence of ... conducted by ...
- some of which can only be thought of as random effects unless we
  - expand beyond sentence-level processing
  - incorporate world knowledge
  - model pragmatics

Upshot is: there is latent variation, thus let's account for it!

Outline

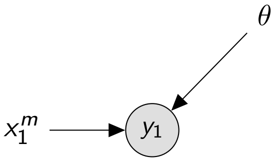
NMT - Recap

A conditional language model with no Markov assumption

x_1^m

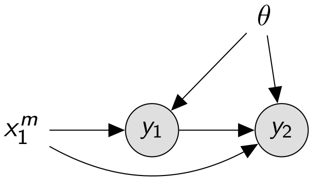
NMT - Recap

A conditional language model with no Markov assumption



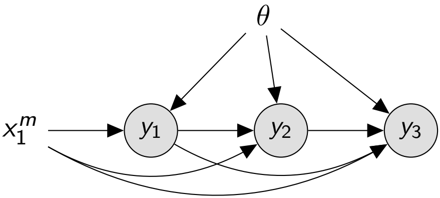
NMT - Recap

A conditional language model with no Markov assumption



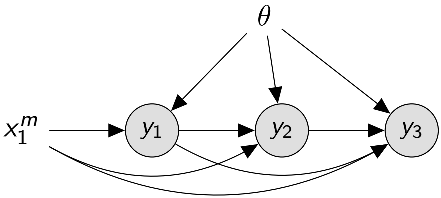
NMT - Recap

A conditional language model with no Markov assumption



NMT - Recap

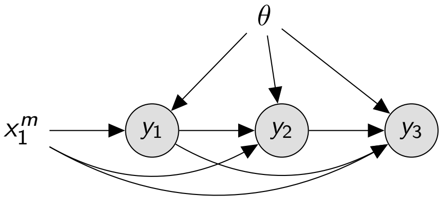
A conditional language model with no Markov assumption



$$P(y_1^n | x_1^m, \theta) = \prod_{i=1}^{n} P(y_i | x_1^m, y_{<i}, \theta)$$

NMT - Recap

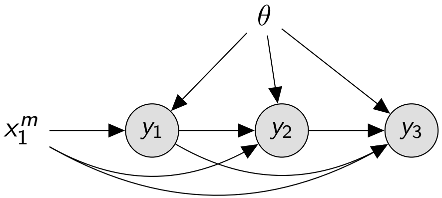
A conditional language model with no Markov assumption



$$P(y_1^n | x_1^m, \theta) = \prod_{i=1}^{n} P(y_i | x_1^m, y_{<i}, \theta)$$
$$= \prod_{i=1}^{n} \text{Cat}(y_i | f_\theta(x_1^m, y_{<i}))$$

NMT - Recap

A conditional language model with no Markov assumption



$$P(y_1^n|x_1^m, \theta) = \prod_{i=1}^{n} P(y_i|x_1^m, y_{<i}, \theta)$$
$$= \prod_{i=1}^{n} \mathrm{Cat}(y_i|f_\theta(x_1^m, y_{<i}))$$

$f_\theta(\cdot)$ is computed by NN architecture with softmax output

NMT - Output

We use a NN to predict a distribution over target words conditioned on source and target prefix

NMT - Output

We use a NN to predict a distribution over target words conditioned on source and target prefix

The factorisation makes no Markov assumptions

NMT - Output

We use a NN to predict a distribution over target words
conditioned on source and target prefix

The factorisation makes no Markov assumptions
but NMT still outputs a single distribution

$$P(y_1^n|x_1^m, \theta) = \prod_{i=1}^{n} \mathrm{Cat}(y_i|f_\theta(x_1^m, y_{<i}))$$

NMT - Output

We use a NN to predict a distribution over target words conditioned on source and target prefix

The factorisation makes no Markov assumptions but NMT still outputs a single distribution

$$P(y_1^n|x_1^m, \theta) = \prod_{i=1}^{n} \text{Cat}(y_i|f_\theta(x_1^m, y_{<i}))$$

Isn't it reasonable to expect that the data is a mixture of various distributions?

NMT - Output

We use a NN to predict a distribution over target words conditioned on source and target prefix

The factorisation makes no Markov assumptions but NMT still outputs a single distribution

$$P(y_1^n|x_1^m, \theta) = \prod_{i=1}^{n} \mathrm{Cat}(y_i|f_\theta(x_1^m, y_{<i}))$$

Isn't it reasonable to expect that the data is a mixture of various distributions?

- $P(y_i|x_1^m, y_{<i}, \text{voice})$

NMT - Output

We use a NN to predict a distribution over target words conditioned on source and target prefix

The factorisation makes no Markov assumptions but NMT still outputs a single distribution

$$P(y_1^n|x_1^m, \theta) = \prod_{i=1}^{n} \mathrm{Cat}(y_i|f_\theta(x_1^m, y_{<i}))$$

Isn't it reasonable to expect that the data is a mixture of various distributions?

- $P(y_i|x_1^m, y_{<i}, \text{voice})$
- $P(y_i|x_1^m, y_{<i}, \text{tired/rested})$

NMT - Output

We use a NN to predict a distribution over target words conditioned on source and target prefix

The factorisation makes no Markov assumptions but NMT still outputs a single distribution

$$P(y_1^n|x_1^m, \theta) = \prod_{i=1}^{n} \mathrm{Cat}(y_i|f_\theta(x_1^m, y_{<i}))$$

Isn't it reasonable to expect that the data is a mixture of various distributions?

- $P(y_i|x_1^m, y_{<i}, \text{voice})$
- $P(y_i|x_1^m, y_{<i}, \text{tired}/\text{rested})$
- $P(y_i|x_1^m, y_{<i}, \text{Br}/\text{Pt})$

Outline

Idea

Account for variation through latent variables on target positions

Idea

Account for variation through latent variables on target positions

Motivations

- capture linguistic phenomena

Idea

Account for variation through latent variables on target positions

Motivations

- capture linguistic phenomena
- better BLEU score

Idea

Account for variation through latent variables on target positions

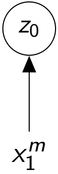
Motivations

- capture linguistic phenomena
- better BLEU score
- we know variation exists, so let's model it

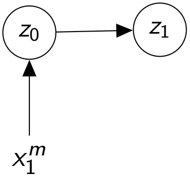
Stochastic Decoder Model

x_1^m

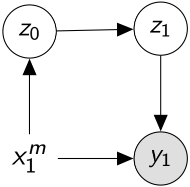
Stochastic Decoder Model



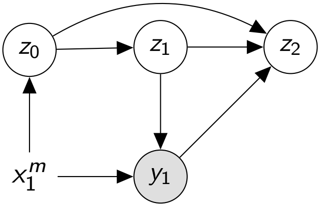
Stochastic Decoder Model



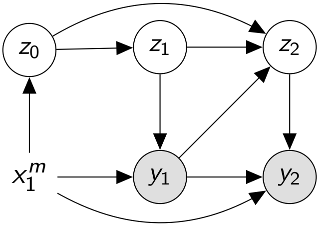
Stochastic Decoder Model



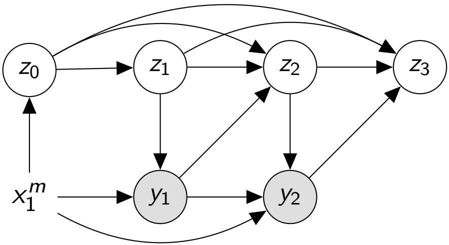
Stochastic Decoder Model



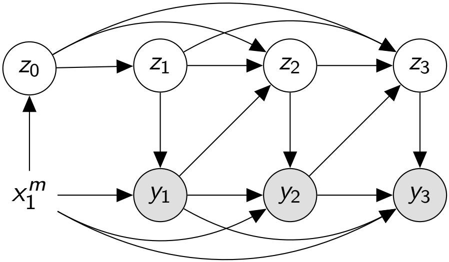
Stochastic Decoder Model



Stochastic Decoder Model



Stochastic Decoder Model



Stochastic Decoder Model

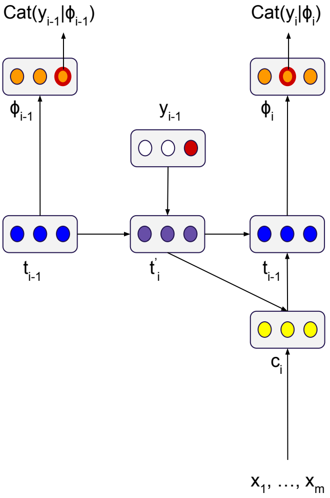
Joint distribution

$$P(y_1^n, z_0^n | x_1^m) = \mathcal{N}(z_0 | \mu_0, \sigma_0^2) \prod_{i=1}^{n} \mathcal{N}(z_i | \mu_i, \sigma_i^2) \times \mathrm{Cat}(y_i | \phi_i)$$

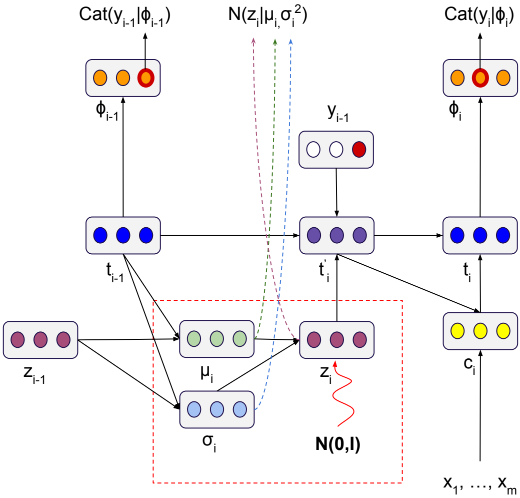
- Gaussian latent variables
  location and scale computed by NN architectures
- Categorical observations
  word probabilities computed by NN architectures

Architecture

Deterministic



Stochastic



Training

Marginalisation is not tractable

$$P(y_1^n|x_1^m) = \int P(y_1^n, z_0^n|x_1^m)\,\mathrm{d}z_0^n$$

Training

Marginalisation is not tractable

$$
\begin{aligned}
P(y_1^n|x_1^m) &= \int P(y_1^n, z_0^n|x_1^m)\,\mathrm{d}z_0^n \\
&= \int \mathcal{N}(z_0|\mu_0, \sigma_0^2) \prod_{i=1}^{n} \mathcal{N}(z_i|\mu_i, \sigma_i^2) \times \mathrm{Cat}(y_i|\phi_i)\,\mathrm{d}z_0^n
\end{aligned}
$$

Training

Marginalisation is not tractable

$$
\begin{aligned}
P(y_1^n|x_1^m) &= \int P(y_1^n, z_0^n|x_1^m)\,\mathrm{d}z_0^n \\
&= \int \mathcal{N}(z_0|\mu_0, \sigma_0^2) \prod_{i=1}^{n} \mathcal{N}(z_i|\mu_i, \sigma_i^2) \times \mathrm{Cat}(y_i|\phi_i)\,\mathrm{d}z_0^n
\end{aligned}
$$

Express a lowerbound in terms of an auxiliary model

$$
\log P(y_1^n|x_1^m) \geq \underbrace{\mathbb{E}_{q(z_0^n)}\left[\log P(y_1^n, z_0^n|x_1^m)\right] + \mathbb{H}(q(z_0^n))}_{\text{ELBO}}
$$

Designing auxiliary model

Considerations

- easy to sample from
- easy to evaluate at a point
- reparameterisable
  $\mathbb{E}_{q(z|\lambda)}[\psi(z)] = \mathbb{E}_{q(\epsilon)}[\psi(z = h^{-1}(\lambda, \epsilon))]$
- preferably an exponential family

Inference model

Amortised inference

$$q(z_0^n) = q(z_0|x_1^m, y_1^n) \prod_{i=1}^{n} q(z_i|x_1^m, y_1^n, z_{<i})$$

Inference model

Amortised inference

$$q(z_0^n) = q(z_0|x_1^m, y_1^n) \prod_{i=1}^{n} q(z_i|x_1^m, y_1^n, z_{<i})$$
$$= \mathcal{N}(z_0|u_0, s_0^2) \prod_{i=1}^{n} \mathcal{N}(z_i|u_i, s_i^2)$$

- no Markov assumption (this is not mean field VI)

Inference model

Amortised inference

$$\begin{aligned} q(z_0^n) &= q(z_0|x_1^m, y_1^n) \prod_{i=1}^{n} q(z_i|x_1^m, y_1^n, z_{<i}) \\ &= \mathcal{N}(z_0|u_0, s_0^2) \prod_{i=1}^{n} \mathcal{N}(z_i|u_i, s_i^2) \end{aligned}$$

- no Markov assumption (this is not mean field VI)
- they condition on x_1^m and y_1^n

Inference model

Amortised inference

$$
\begin{aligned}
q(z_0^n) &= q(z_0|x_1^m, y_1^n) \prod_{i=1}^{n} q(z_i|x_1^m, y_1^n, z_{<i}) \\
&= \mathcal{N}(z_0|u_0, s_0^2) \prod_{i=1}^{n} \mathcal{N}(z_i|u_i, s_i^2)
\end{aligned}
$$

- no Markov assumption (this is not mean field VI)
- they condition on x_1^m and y_1^n
- parameters of variational factors are predicted by NN architectures

Cascade of ELBOs

$$\mathrm{ELBO} = \mathrm{ELBO}_0 + \mathbb{E}\left[\mathrm{ELBO}_1 + \mathbb{E}\left[\mathrm{ELBO}_2 + \ldots\right]\right]$$

Cascade of ELBOs

$$\text{ELBO} = \text{ELBO}_0 + \mathbb{E}\left[\text{ELBO}_1 + \mathbb{E}\left[\text{ELBO}_2 + \ldots\right]\right]$$

ELBO_i is

$$\underbrace{\mathbb{E}_{\frac{z_i - u_i}{\sigma_i} \sim \mathcal{N}(0,I)}\left[\log \text{Cat}(y_i|\phi_i)\right]}_{\text{"reconstruction term"}} - \underbrace{\text{KL}\left(\mathcal{N}(z_i|u_i, s_i^2) \mid\mid \mathcal{N}(z_i|\mu_i, \sigma_i^2)\right)}_{\text{"complexity cost"}}$$

Cascade of ELBOs

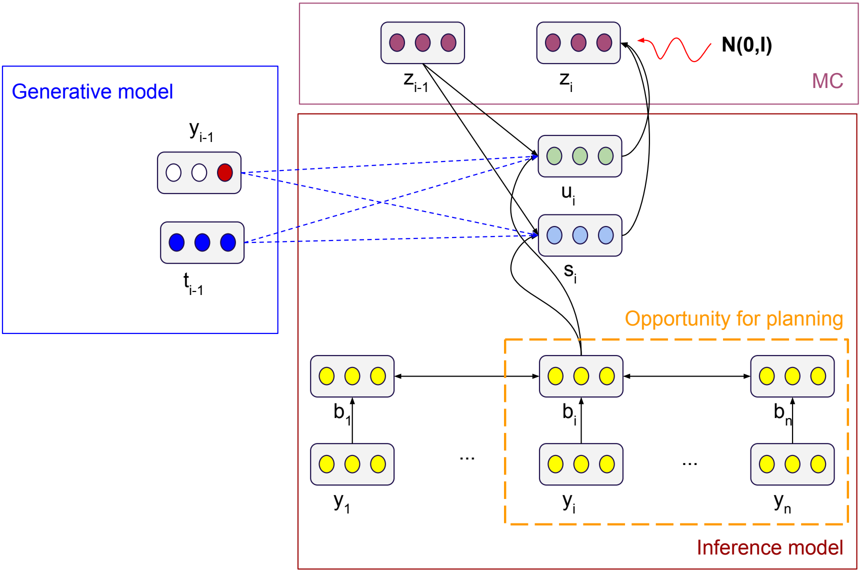
$$\mathrm{ELBO} = \mathrm{ELBO}_0 + \mathbb{E}\left[\mathrm{ELBO}_1 + \mathbb{E}\left[\mathrm{ELBO}_2 + \ldots\right]\right]$$

ELBO_i is

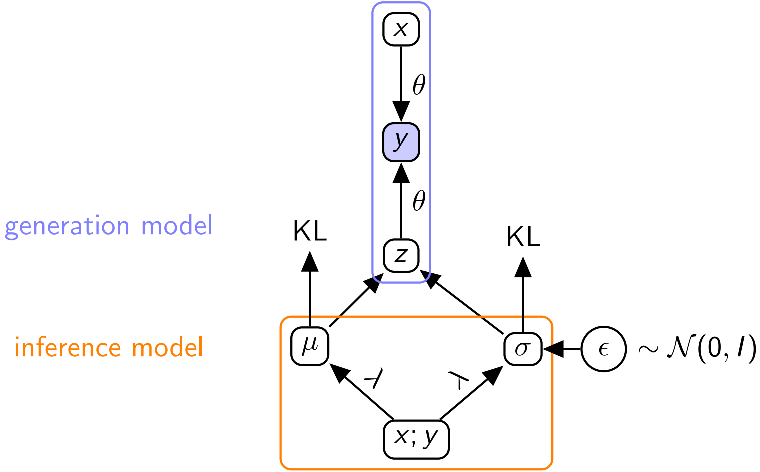
$$\underbrace{\mathbb{E}_{\frac{z_i - u_i}{\sigma_i} \sim \mathcal{N}(0,I)}\left[\log \mathrm{Cat}(y_i|\phi_i)\right]}_{\text{“reconstruction term”}} - \underbrace{\mathrm{KL}\left(\mathcal{N}(z_i|u_i, s_i^2)\;||\;\mathcal{N}(z_i|\mu_i, \sigma_i^2)\right)}_{\text{“complexity cost”}}$$

- reparameterisation enables backpropagation through samples
- KL is analytical for Gaussians
- computing the ELBO is a sequential process

Inference network



Computation graph



Outline

Data

IWSLT 2016

Data	Arabic	Czech	French	German
Train	224,125	114,389	220,399	196,883
Dev	6,746	5,326	5,937	6,996
Test	2,762	2,762	2,762	2,762

Table : Number of sentence pairs

Systems

Bahdanau et al. (2014) baseline

- BiLSTM encoder
- attention
- LSTM decoder

Zhang et al. (2016) SENT

- $P(y_1^n, z_0|x_1^m) = p(z_0|x_1^m)P(y_1^n|x_1^m, z_0)$

Stochastic decoder SDEC

- Code and workflow:
  `https://github.com/philschulz/stochastic-decoder`
- Paper: ACL2018
  joint work with Philip Schulz and Trevor Cohn

Hyperparameters

- vocab size: 50,000 sentence pieces
- framework: Sockey
- 1028 LSTM units (512 for each LSTM encoder)
- 256 units for attention
- Adam 10^{-3}
- Dropout: 0.5 (based on dev BLEU of baseline)
- KL scaling: 0 to 1 with steps $20,000^{-1}$
- Test decoding: beam size 5, latent variables deterministically set to the mean

Results

Model	Dropout	LatentDim	Arabic	Czech	French	German
Sockeye	None	None	8.2	6.9	23.5	14.3
Sockeye	0.5	None	8.4	7.4	24.4	15.1
SENT	0.5	64	8.4	7.3	24.8	15.3
SENT	0.5	128	8.7	7.4	24.0	15.7
SENT	0.5	256	8.9	7.4	24.7	15.5
SDEC	0.5	64	8.2	7.7	25.3	15.4
SDEC	0.5	128	8.8	7.5	24.2	15.6
SDEC	0.5	256	8.7	7.5	23.2	15.9

Table : BLEU

Examples

Source	Coincidentally, at the same time, the first easy-to-use clinical tests for diagnosing autism were introduced.
SENT	Im gleichen Zeitraum wurden die ersten einfachen klinischen Tests für Diagnose getestet.
SDEC	Übrigens, zur gleichen Zeit, wurden die ersten einfache klinische Tests für die Diagnose von Autismus eingeführt.
SDEC	Übrigens, zur gleichen Zeit, <u>waren</u> die ersten einfache klinische Tests für die Diagnose von Autismus eingeführt <u>worden</u>.

Figure : The example shows alternation between the German simple past and past perfect. The past perfect introduces a long range dependency between the main and auxiliary verb (underlined) that the model handles well.

Outline

Related work

- Bayer and Osendorfer (2014) noise sources have no sequential dependencies

Related work

- Bayer and Osendorfer (2014) noise sources have no sequential dependencies
- Chung et al. (2015) stochastic RNN, but no lookahead essentially limited to be as effective as the prior

Related work

- Bayer and Osendorfer (2014) noise sources have no sequential dependencies
- Chung et al. (2015) stochastic RNN, but no lookahead essentially limited to be as effective as the prior
- Fraccaro et al. (2016) two separate RNNs: one stochastic, one deterministic

Related work

- Bayer and Osendorfer (2014) noise sources have no sequential dependencies
- Chung et al. (2015) stochastic RNN, but no lookahead essentially limited to be as effective as the prior
- Fraccaro et al. (2016) two separate RNNs: one stochastic, one deterministic
- Su et al. (2018) stochastic RNN for NMT, but no lookahead

Is there a formal argument in favour of this model?

Is there a formal argument in favour of this model?

NMT makes no Markov assumptions

$$P(y_i|x_1^m, y_{<i}) = \exp\left(\eta(x_1^m, y_{<i})^\top t(y_i) - a(\eta)\right)$$

Is there a formal argument in favour of this model?

NMT makes no Markov assumptions

$$P(y_i|x_1^m, y_{<i}) = \exp\left(\eta(x_1^m, y_{<i})^\top t(y_i) - a(\eta)\right)$$

but it makes a **statistical assumption**

Is there a formal argument in favour of this model?

NMT makes no Markov assumptions

$$P(y_i|x_1^m, y_{<i}) = \exp\left(\eta(x_1^m, y_{<i})^\top t(y_i) - a(\eta)\right)$$

but it makes a **statistical assumption**

- data distribution is an exponential family
- NN controls its natural parameter

Is there a formal argument in favour of this model?

NMT makes no Markov assumptions

$$P(y_i|x_1^m, y_{<i}) = \exp\left(\eta(x_1^m, y_{<i})^\top t(y_i) - a(\eta)\right)$$

but it makes a **statistical assumption**

- data distribution is an exponential family
- NN controls its natural parameter

What if the data distribution is not an exponential family?

Is there a formal argument in favour of this model?

NMT makes no Markov assumptions

$$P(y_i|x_1^m, y_{<i}) = \exp\left(\eta(x_1^m, y_{<i})^\top t(y_i) - a(\eta)\right)$$

but it makes a **statistical assumption**

- data distribution is an exponential family
- NN controls its natural parameter

What if the data distribution is not an exponential family?

Stochastic decoder

$$P(y_i|x_1^m, y_{<i}) = \int p(z_0^i) \exp\left(\eta(x_1^m, y_{<i}, z_0^i)^\top t(y_i) - a(\eta)\right) \mathrm{d}z_0^i$$

is more general than an exponential family

Remarks

Criticism

- systematically quantify variation in translation data
- systematically diagnose latent space

Remarks

Criticism

- systematically quantify variation in translation data
- systematically diagnose latent space

Ongoing and beyond

- fully correlated Gaussians
- non-Gaussian variables
- latent feature model

Message

DL helps us get rid of unrealistic modelling assumptions
and that's great

- finite memory for LMs
- 1-1 alignments for MT

Message

DL helps us get rid of unrealistic modelling assumptions
and that's great

- finite memory for LMs
- 1-1 alignments for MT

but statistical assumptions may also correspond to inductive bias

- better reflect the nature of the data
- better encode expert knowledge about the problem

Literature I

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. 2014. URL `http://arxiv.org/abs/1409.0473`.

Justin Bayer and Christian Osendorfer. Learning stochastic recurrent networks. *arXiv preprint arXiv:1411.7610*, 2014.

Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, 39(3):463–472, 2013.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015.

Literature II

Marco Fraccaro, Søren Kaae Sø nderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2199–2207. 2016. URL http://papers.nips.cc/paper/6039-sequential-neural-models-with-stochastic-layers.pdf.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. Variational recurrent neural machine translation. *arXiv preprint arXiv:1801.05119*, 2018.

Literature III

Biao Zhang, Deyi Xiong, jinsong su, Hong Duan, and Min Zhang. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas, November 2016. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1050`.