Modeling Latent Sentence Structure in Neural Machine Translation

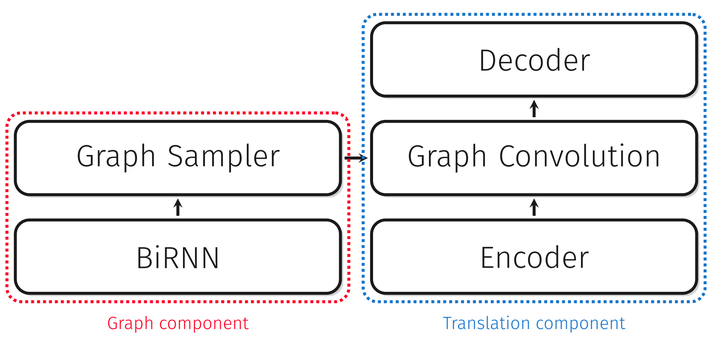
Joost Bastings[1], Wilker Aziz[1], Ivan Titov[1,2], Khalil Sima'an[1]

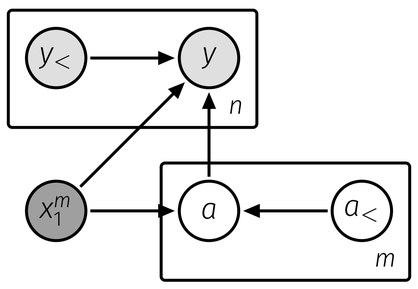
[1]University of Amsterdam, [2]University of Edinburgh

Model

We present a **deep generative model**; a probabilistic model whose two components are parameterized by neural nets:

1. a **graph component** that stochastically samples a latent graph a_1^m conditioned on the source sentence x_1^m
2. a **translation component** that conditions on source sentence x_1^m and sampled graph a_1^m to predict the target sentence y_1^n



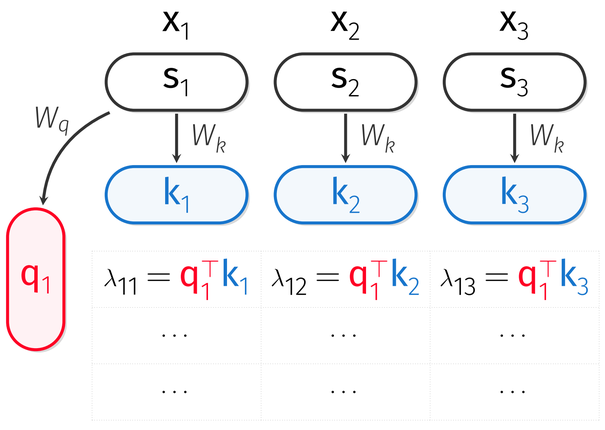


Graph Component

- **Samples** for each source position i an m-dimensional **probability vector** A_i:

$$A_{i \mid} x_1^m \sim \text{Concrete}(\tau, \boldsymbol{\lambda}_i)$$

- We interpret A_i as a head distribution
- a_{ik} is the **relative strength** of the edge $x_i \to x_k$
- 'Head potentials' $\boldsymbol{\lambda}_i$ are computed with **self-attention**:



Translation Component

- Attentive **encoder-decoder** that incorporates graph a_1^m using **graph convolution** *(see next)*
- Samples a target word at each time step j:

$$Y_j | x_1^m, a_1^m, y_{<j} \sim \text{Cat}(\boldsymbol{\pi}_j)$$
$$\boldsymbol{\pi}_j = f_\theta(x_1^m, a_1^m, y_{<j})$$

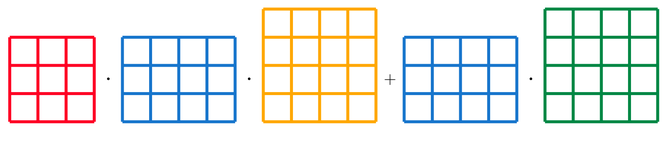
- We experiment with **three encoders** to get $\mathbf{s}_1^m$: embeddings (+PE), CNNs (+PE) and RNNs

Summary

- We incorporate **sentence structure** as a **latent variable** in an encoder-decoder
- We induce it in such a way as to benefit the translation task
- Previous work [1, 2, 3] showed that **syntactic structure** can be beneficial for NMT, but often relies on **supervised parsers**
- The graphs may capture useful dependencies (as evidenced by BLEU), but have difficulty enhancing RNNs

Graph Convolution (GCN)

- We employ **graph convolutional networks** [1] to incorporate graph $\mathbf{A} = a_1^m$ into source word representations $\mathbf{S} = \mathbf{s}_1^m$:



$$\text{GCN}(\mathbf{S}, \mathbf{A}) = \text{ReLU}\left(\mathbf{A}\mathbf{S}\mathbf{W}_{\text{IN}} + \mathbf{S}\mathbf{W}_{\text{LOOP}} + \mathbf{b}\right)$$

- We apply gates to $\mathbf{A}$ and the self-loop

Parameter estimation

- We optimize the following **objective**:

$$\log p(y \mid x) \geq \mathbb{E}\Big[\log p(y \mid x, a)\Big]$$
$$\approx \log p(y \mid x, a)$$
$$a \sim p(a \mid x_1^m)$$

Experiments

- De-En (IWSLT14) & Ja-En (ASPEC)
- Code based on Tensorflow NMT

	Train	Dev	Test	Vocabularies
De-En	153K	7282	6750	32010/22823
Ja-En	2M	1790	1812	16384 (SPM)

Results

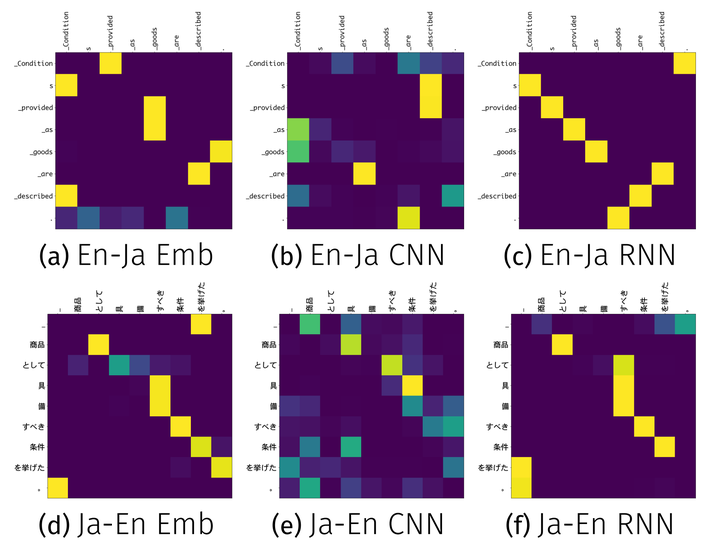
		IWSLT14		WAT17	
	Encoder	De-En	En-De	Ja-En	En-Ja
Ext. baseline	RNN	27.6	-	-	28.5
Baseline	Emb.	22.7	17.9	18.1	18.1
Baseline	CNN	23.6	19.1	23.0	24.6
Baseline	RNN	27.6	22.4	26.0	28.7
Latent Graph	Emb.	24.0	18.7	23.2	24.3
Latent Graph	CNN	24.6	20.3	24.6	26.7
Latent Graph	RNN	27.2	22.4	26.0	29.1

Analysis

	Mean head distance		Mean entropy	
Encoder	Ja-En	En-Ja	Ja-En	En-Ja
Emb.	4.0 ±6.9	3.8 ±5.6	0.49 ±0.18	0.42 ±0.18
CNN	6.1 ±6.5	6.7 ±7.1	1.21 ±0.28	1.47 ±0.30
RNN	4.3 ±6.5	2.0 ±5.4	0.51 ±0.20	0.00 ±0.01

- The **average head distance** shows that the graphs may capture non-local dependencies
- The **average entropy** shows that the graphs have a high degree of sparsity

Examples



(a) En-Ja Emb (b) En-Ja CNN (c) En-Ja RNN

(d) Ja-En Emb (e) Ja-En CNN (f) Ja-En RNN

Conclusions

- We presented a model with separate **graph induction** and **translation** components
- The graphs may capture useful dependencies (as evidenced by BLEU), but have difficulty enhancing RNNs

References

[1] Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. "Graph Convolutional Encoders for Syntax-aware Neural Machine Translation". In: *EMNLP*. 2017.

[2] Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. "Learning to Parse and Translate Improves Neural Machine Translation". In: *ACL*. 2017.

[3] Kazuma Hashimoto and Yoshimasa Tsuruoka. "Neural Machine Translation with Source-Side Latent Graph Parsing". In: *EMNLP*. 2017.

[4] Ke Tran and Yonatan Bisk. "Inducing Grammars with and for Neural Machine Translation". In: *ACL NMT*. 2018.

[5] Adina Williams, Andrew Drozdov, and Samuel R. Bowman. "Do latent tree learning models identify meaningful structure in sentences?". In: *TACL* (2018).

This work was supported by the European Research Council (ERC StG BroadSem 678254) and the Dutch National Science Foundation (NWO VIDI 639.022.518, NWO VICI 277-89-002).