基于协同关系的非自回归生成研究

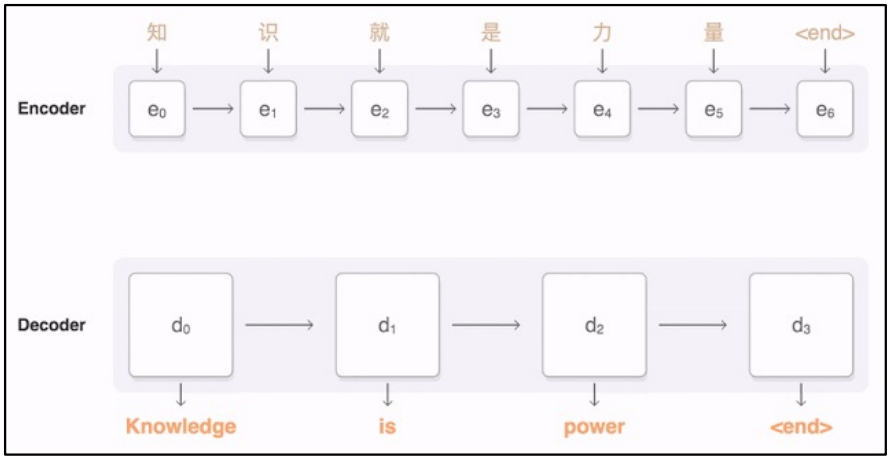
黄书剑 博士 副教授
huangsj@nju.edu.cn
南京大学计算机科学与技术系
计算机软件新技术国家重点实验室

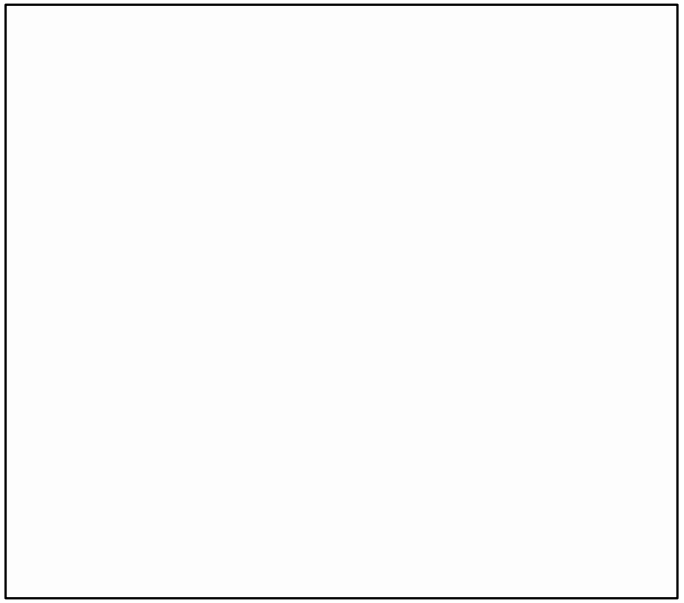
机器翻译的发展

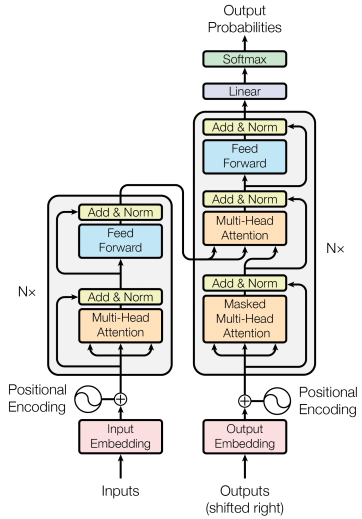
- 借助计算机程序将给定的输入由一种自然语言（源语言）翻译成为另一种自然语言（目标语言）

神经网络机器翻译

- **编码过程的并行化：从 RNN 到 Transformer**
  - 自左向右的时序关系被并行的self-attention计算取代





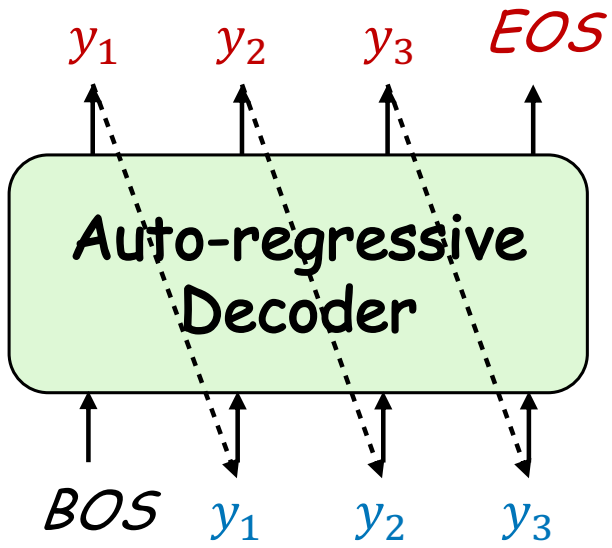


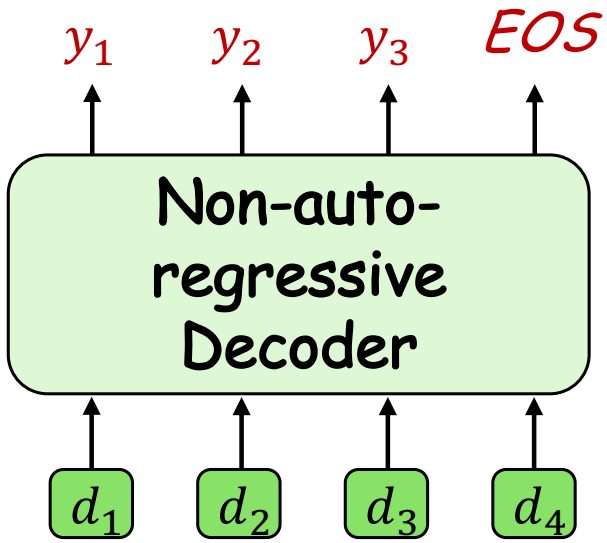
Vaswani et al., 2017

Images from: https://medium.com/analytics-vidhya/transformer-vs-rnn-and-cnn-18eeefa3602b

解码过程的并行化

- 自左向右解码（auto-regressive）v.s. 并行解码（non-auto-regressive）





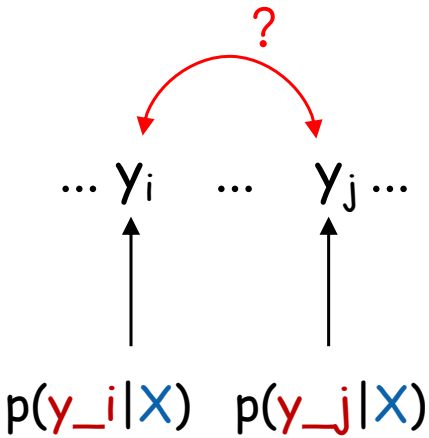
$P(Y|X) = p(y\_1|X) \cdot p(y\_2|y\_1,X) \cdot \cdots \cdot p(y\_n|y\_1,y\_2,\cdots,y\_\{n-1\},X)$

$P(Y|X) \approx p(y\_1|X) \cdot p(y\_2|X) \cdot \cdots \cdot p(y\_n|X)$

Gu et al. 2017

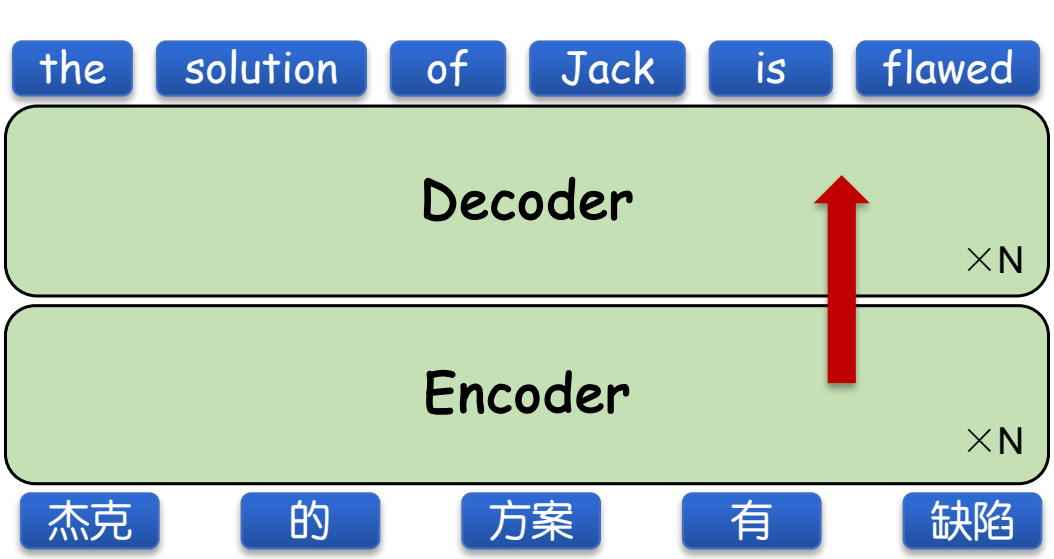
NAT v.s. AT

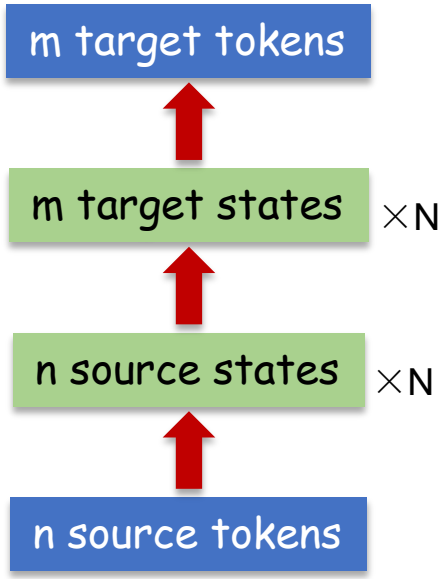
- **优势：**
  - 生成的并行性有助于提升翻译效率
  - 可以减少顺序翻译过程中的错误累积
- **劣势：**
  - 过强的独立性假设可能影响翻译质量



- **生成过程中，文本内部存在怎样的相互关系?**

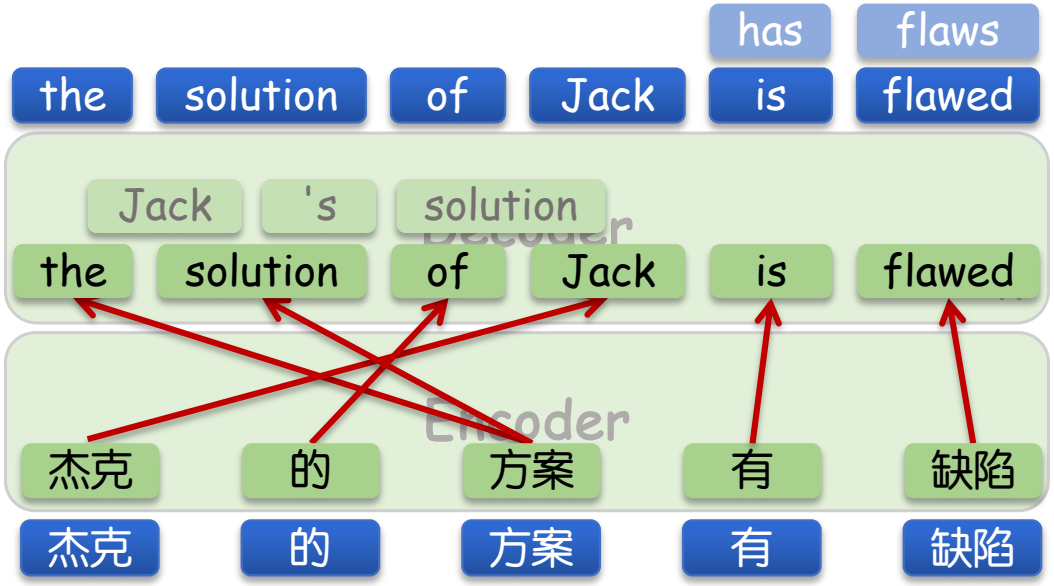
文本之间的协同关系

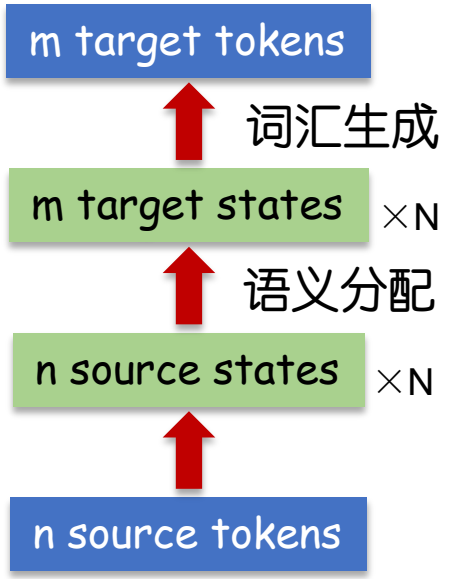

the
solution
of
Jack
is
flawed
Decoder
×N
Encoder
×N
杰克
的
方案
有
缺陷


m target tokens
m target states
×N
n source states
×N
n source tokens

文本之间的协同关系

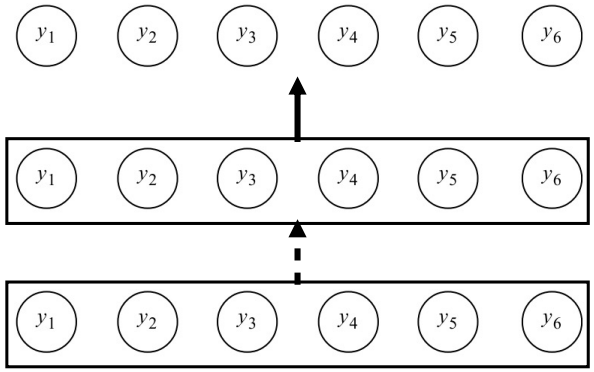
- 语义分配：决定target states（顺序、覆盖）
- 词汇生成：决定target token（一致、关联）





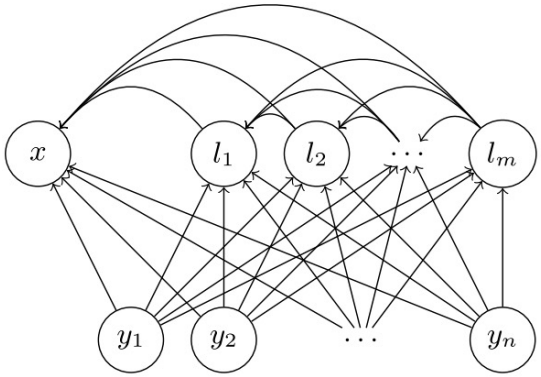
处理文本之间的协同关系的方案

Iterative NAT
(Lee et al. 2018)



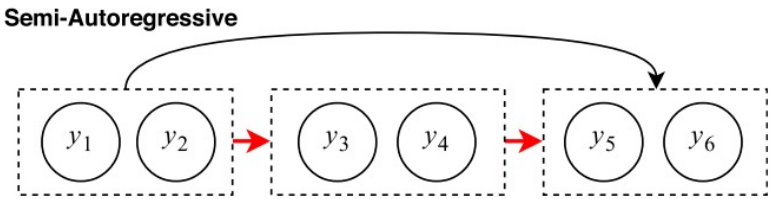
需要多次解码

Latent NAT
(Kaiser et al. 2018)



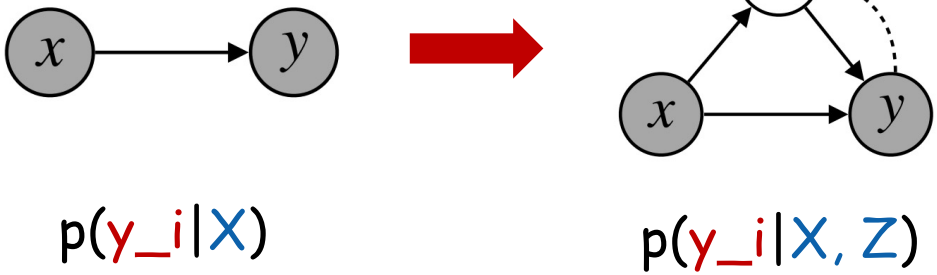
隐变量多为AT

Semi-AT
(Wang et al. 2018)

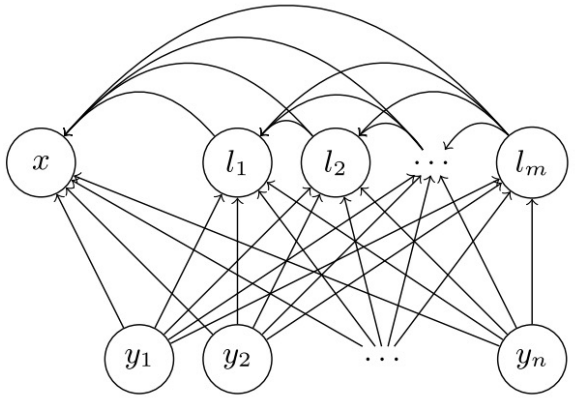


划分难以把握

基于隐变量的方案



p(y_i|X)　　p(y_i|X, Z)



- **将隐变量作为生成的中介，承担协同的功能**
  - 保持Y的独立性，弱化独立性假设
- **潜在优势：相对于生成Y，生成Z的过程可以相对高效**
- **潜在困难：隐变量的学习过程具有挑战**

提纲

- **通过隐变量建模离散类别信息及其关联**
  - Non-Autoregressive Translation by Learning Target Categorical Codes. (NAACL2021)
- **通过位置关系预测建模单词之间的关系**
  - Non-Autoregressive Transformer by Position Learning (arXiv: 1911.10677)
- **引入依存句法结构建模目标端的关联关系**
  - Modeling Target-side Interrelation for Non-Autoregressive Neural Machine Translation (in progress)

合作者：

- 鲍宇，王东琪，戴新宇，陈家骏（南京大学）
- 周浩，李磊，王明轩，封江涛（字节跳动 AI Lab）
- 肖桐（东北大学）

南京大学 NANJING UNIVERSITY

ByteDance AI Lab 字节跳动人工智能实验室

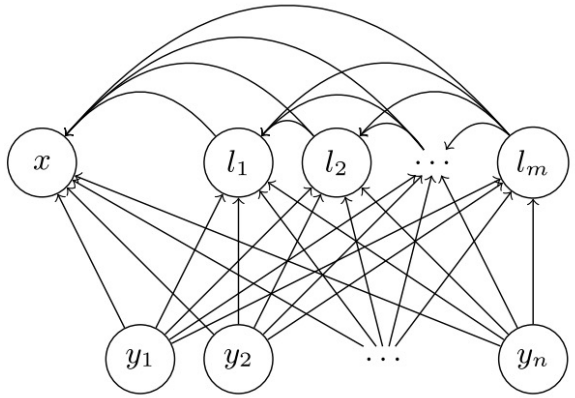
小牛翻译

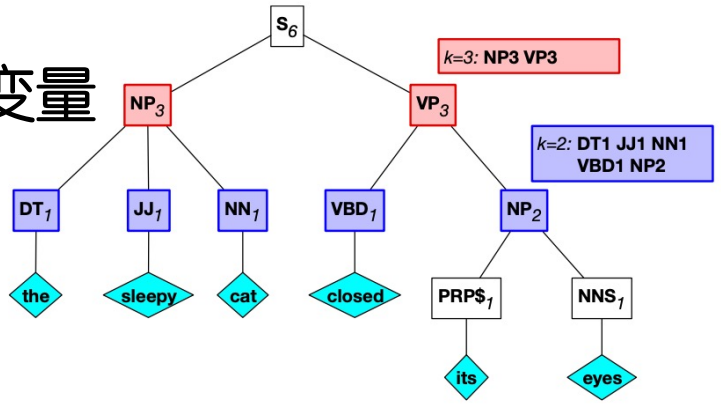
南京大學
NANJING UNIVERSITY

通过隐变量建模离散类别信息及其关联
CATEGORICAL-NAT

如何建模隐变量

- **利用更短的隐变量序列 [Kaiser et al. 2018]**
  - 每个隐变量代表k个连续单词（chuck）
  - 缺陷：
    - latent code > 32k，效率受到限制
    - chuck 难以适应不同情况
- **利用外部指定的序列 [Akoury et al. 2019]**
  - 根据句法树的结构标记和长度决定中间变量
  - 缺陷：
    - 依赖于外部分析结果
    - 不同层次的句法标记带来解码困难

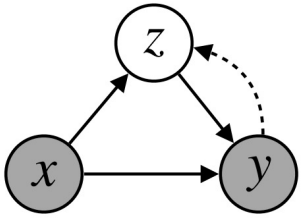




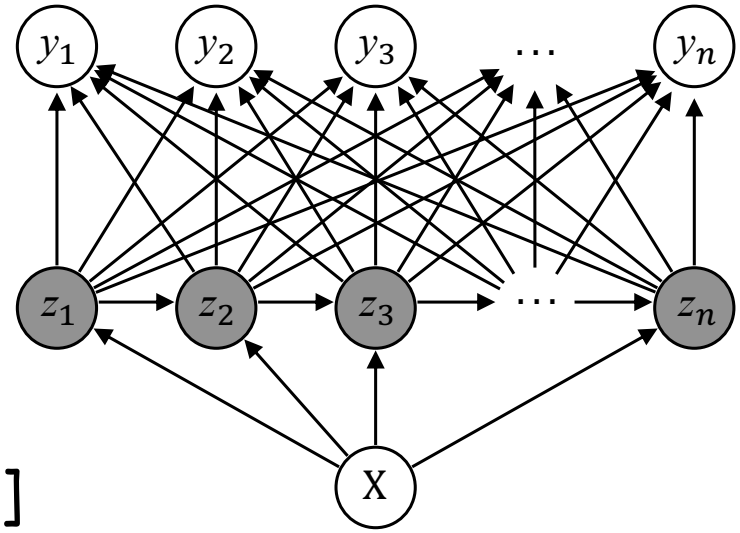
是否能够学到更加有效且高效的隐变量用于协同?

我们的方案（CNAT）

- **自动学习离散隐变量作为关联**
  - VQ-VAE + EMA [Kaiser et al., 2018]



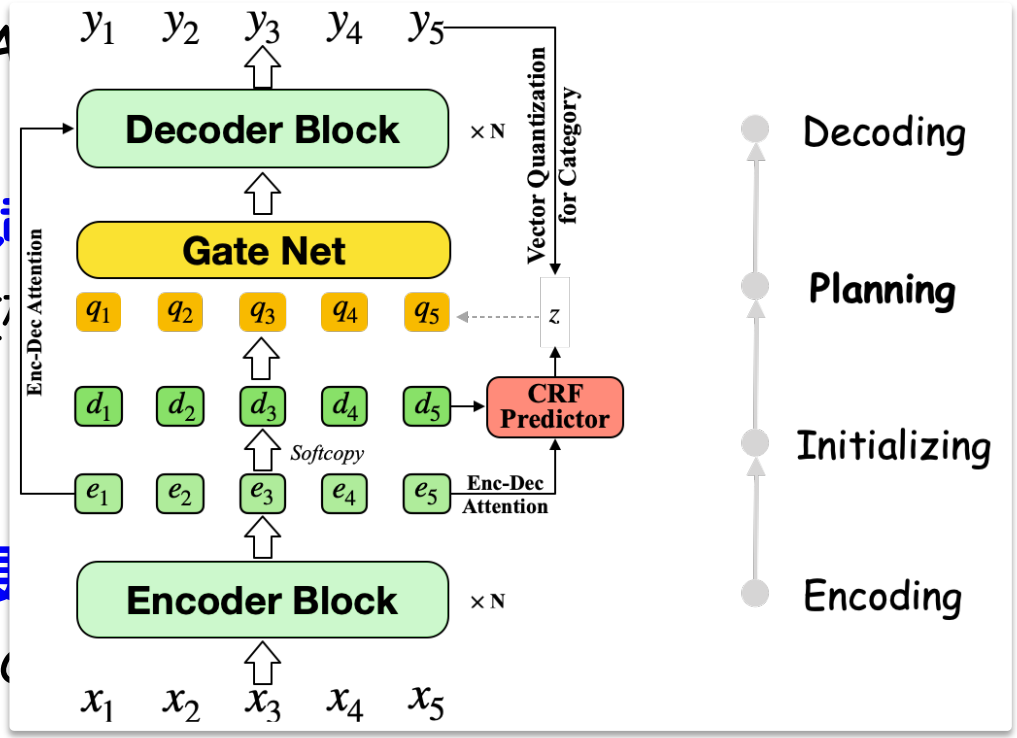
- **为每个单词学习独立的隐变量**
  - 避免建模变长的序列
  - latent code 更加简单 n=64



- **可以使用更复杂的技术提升协同关系建模**
  - linear chain CRF [Lafferty et al. 2001]

我们的方案（CNAT）

- 自动学习离散隐变量作为关联
  - VQ-VA
- 为每个单词
  - 避免建
  - latent
- 可以使用更
  - linear



模型训练

- GateNet
  - 将隐变量z的code q和原先的decoder input一同输入解码

$$o_i = d_i * g_i + q_i * (1 - g_i)$$

$$g_i = \sigma(\mathrm{FFN}([d_i; q_i]))$$

- 同时优化协同任务（CRF）和翻译任务（NAT）

$$\mathcal{L}_{\mathrm{crf}} = -\log p(\boldsymbol{z}^{\mathrm{ref}}|\boldsymbol{x})$$

$$\mathcal{L}_{\mathrm{NAT}} = -\log p(\boldsymbol{y}|\boldsymbol{z}^{\mathrm{mix}}, \boldsymbol{x}; \theta)$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{NAT}} + \alpha\mathcal{L}_{\mathrm{crf}}$$

基础实验

- 与基础NAT等模型的比较（without KD）

Model	WMT14 EN-DE	WMT14 DE-EN	IWSLT14 DE-EN
LV-NAR	11.80	/	/
AXE CMLM	20.40	24.90	/
SynST	20.74	25.50	23.82
Flowseq	20.85	25.40	24.75
NAT (ours)	9.80	11.02	17.77
CNAT (ours)	**21.30**	**25.73**	**29.81**

- 极大的提升了NAT的能力，超过了SynST等现有方法

综合实验

- 加入KD（left）及reranking（right）等技术

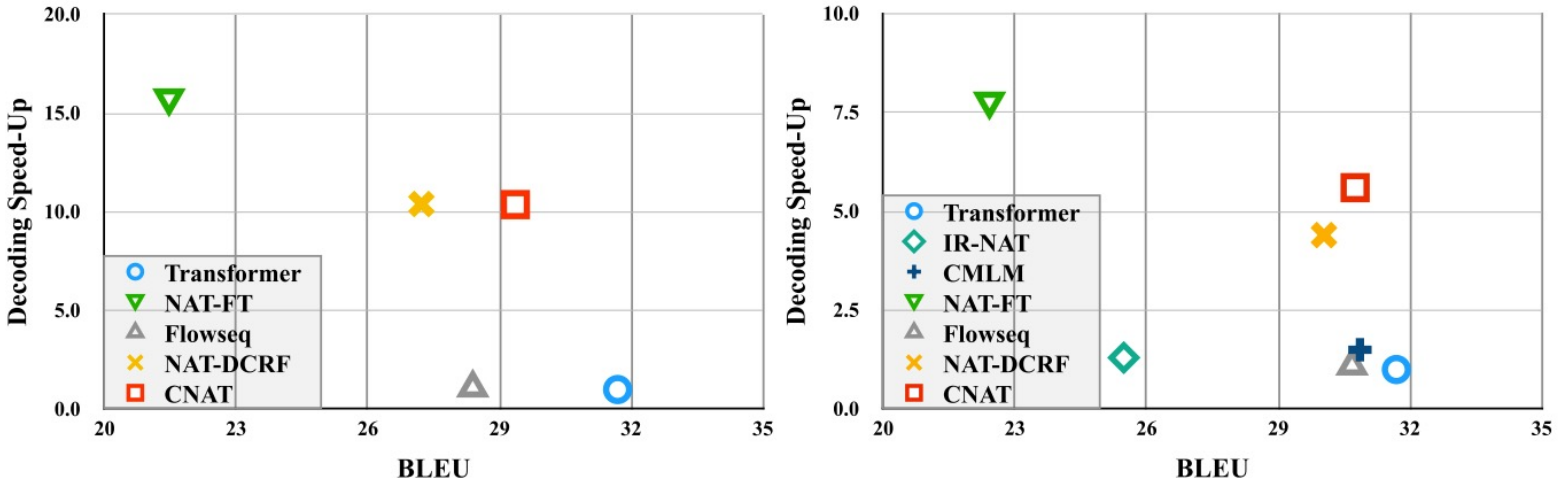
Model	WMT14 EN-DE	WMT14 DE-EN	IWSLT14 DE-EN
NAT-FT	17.69	21.47	/
LT	19.80	/	/
NAT-REG	20.65	24.77	23.89
imitate-NAT	22.44	25.67	/
Flowseq	23.72	28.39	27.55
NAT-DCRF	23.44	27.22	27.44
Transformer (ours)	27.33	31.69	34.29
NAT (ours)	17.69	18.93	23.78
CNAT (ours)	**25.56**	**29.36**	**31.15**

Model	N	WMT14 EN-DE	WMT14 DE-EN
NAT-FT	10	18.66	22.42
NAT-FT	100	19.17	23.20
LT	10	21.00	/
LT	100	22.50	/
NAT-REG	9	24.61	28.90
imitate-NAT	9	24.15	27.28
Flowseq	15	24.70	29.44
Flowseq	30	25.31	30.68
NAT-DCRF	9	26.07	29.68
NAT-DCRF	19	**26.80**	30.04
Transformer (ours)	-	27.33	31.69
CNAT (ours)	9	26.60	**30.75**

- 达到或超越了投稿时的一系列方法

翻译质量 v.s. 解码效率

- CNAT可以更好的平衡翻译质量和解码效率



(a) Pure NAT decoding.

(b) NAT with advanced decoding techniques.

Figure 2: BLEU and decoding speed-up of NAT models on WMT14 DE-EN test set. Each point represents the decoding method run with its corresponding setting in Table 2, Table 3 or Table 4.

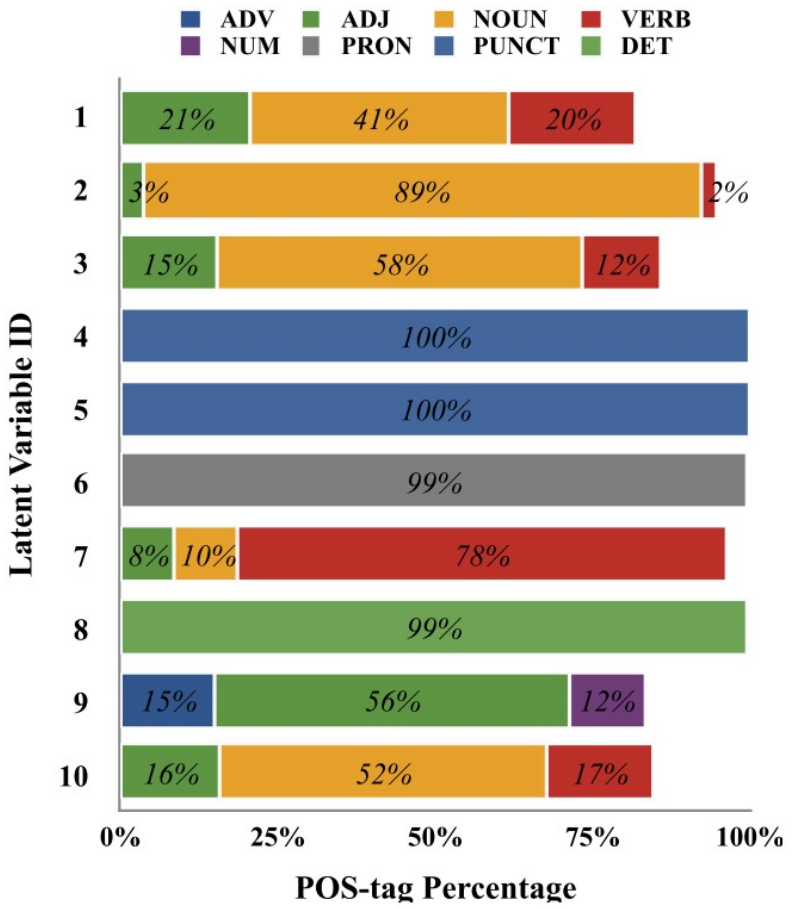
Ablation

- latent code数量 K = 64 即可取得较好的效果 (line 1-3)
- CRF比简单的AR有更好效果(line 2 v.s. line 5, +3 BLEU)
- GateNet有重要效果 (line 2 v.s. line 4, +2 BLEU)

Line	K			Predictor		GateNet	BLEU
	32	64	128	CRF	AR		
1	✓			✓		✓	30.13
2		✓		✓		✓	**31.87**
3			✓	✓		✓	30.82
4		✓		✓			29.32
5		✓			✓	✓	28.23
6		✓			✓		24.00
7			✓		✓		25.43
8							24.25

隐变量的可解释性分析

- **频率最高的10个隐变量的POS分布**
  - 与POStag有较为明显的对应
    - 4、5 PUNCT
    - 6 PRON
    - 8 DET
  - 并不完全与某个POS相同



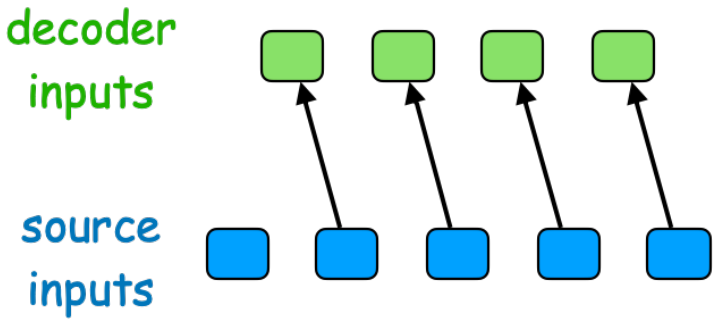
小结

- **可以自动学习到少量离散的隐变量作为单词的代表**
  - 通过避免建模多个词构成的序列
- **由此，可以进行更复杂的单词关联关系建模**
  - 通过line-chain CRF
- **少量离散隐变量具有一定的可解释性**

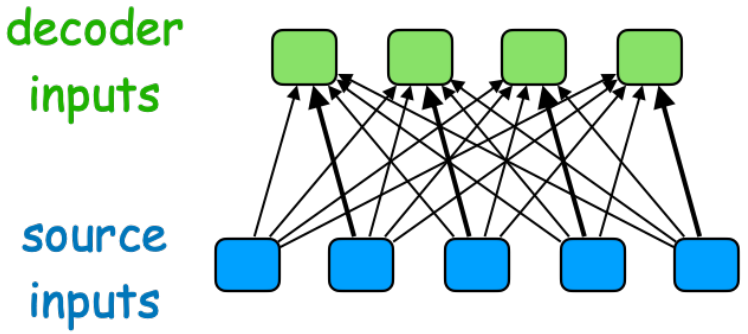
南京大學
NANJING UNIVERSITY

通过位置关系预测建模单词之间的关系
POSITIONAL-NAT

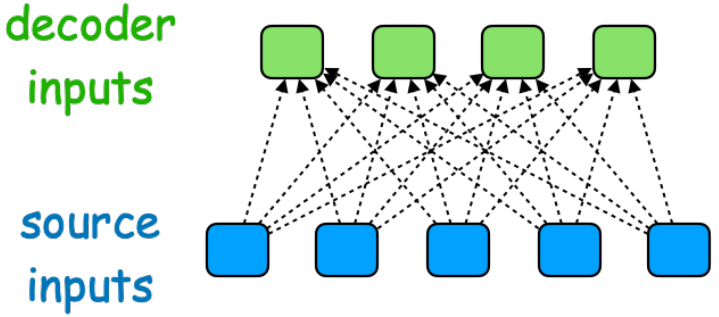
如何确定解码器的状态?



Uniform copy
[Gu et al. 2018]



Softcopy [Wei et al. 2019]
(heuristic interpolation)

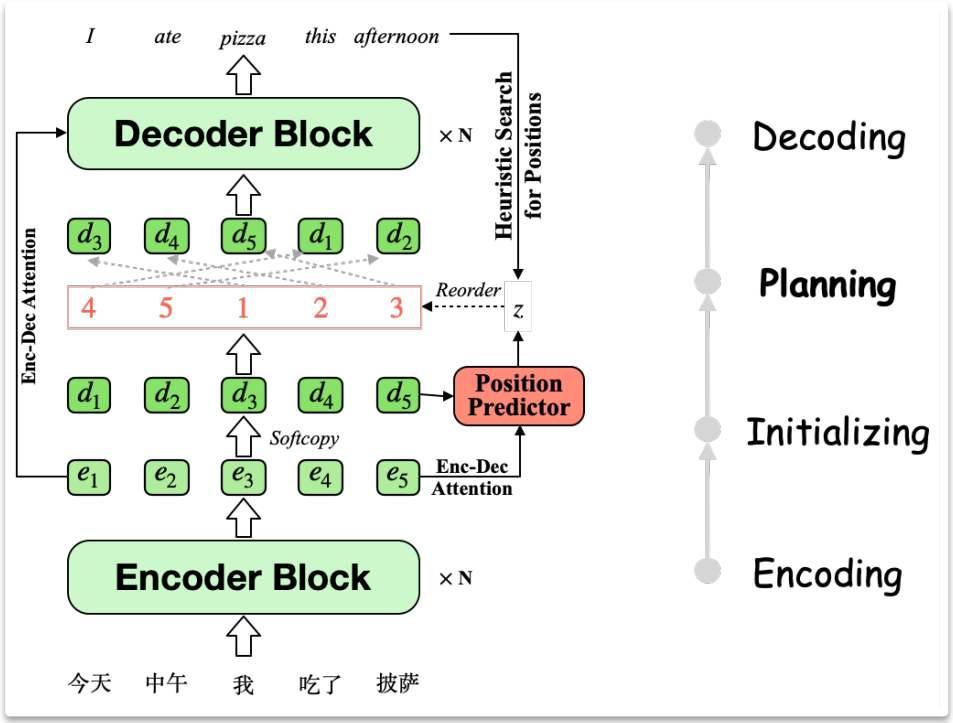


Attention [Qian et al. 2020]
(learnable interpolation)

解码器初始状态仅是给定位置编码器状态的融合，语义分配缺少有效协同

调序能力和调序的协同

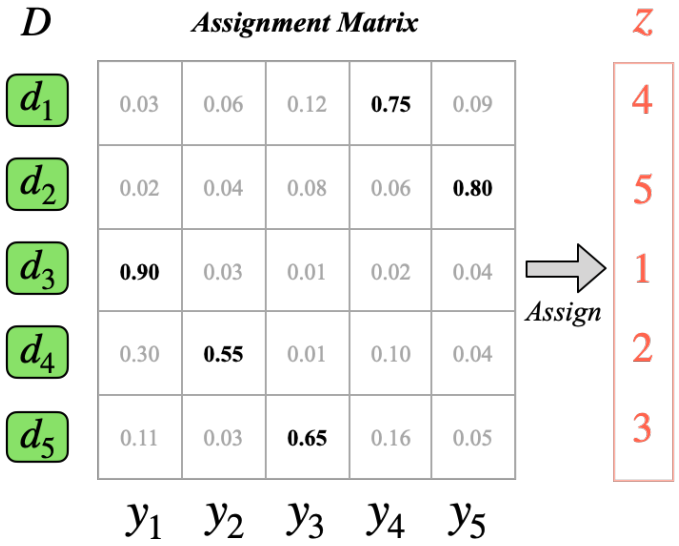
- **语义分配缺少有效协同**
  - 对于顺序差别较大的语言难以有效完成调序
- **解决方案 P-NAT:**
  - 引入位置隐变量
  - 显式调整输出的顺序



调序模型

- **为每个解码器输入d预测其在最终结果中的位置**
  - AT调序：自左向右依次预测每一个d的位置
- **训练：将训练数据中的y和d进行对应**
  - 线性规划问题，采用启发式方法进行求解

$$y_{1:m} \longleftrightarrow d_{1:m}$$



基础实验结果

- 与基础NAT等模型的比较（without KD）

Model	BLEU↑ EN-DE	DE-EN
LV-NAR [26]	11.80	/
CMLM [27]	10.88	/
Flowseq [20]	18.55	23.36
NAT(ours)	9.80	11.02
PNAT(ours)	**19.73 (+9.93)**	**24.04 (+13.02)**

- 极大的提升了NAT的能力

综合实验

- 加入KD（left）及reranking（right）等技术

Model	BLEU↑ EN-DE	BLEU↑ DE-EN
NAT-FT [3]	17.69	21.47
LT [25]	19.80	/
IR-NAR [4]	13.91	16.77
CMLM [27]	15.06	19.26
ENAT [24]	20.65	23.02
NAT-REG [5]	20.65	24.77
imitate-NAT [6]	22.44	25.67
Flowseq [20]	21.45	26.16
Transformer(our)	27.30	31.33
NAT (ours)	17.69	18.93
PNAT (ours)	**23.05 (+5.36)**	**27.18 (+8.25)**

Model	N	BLEU↑ EN-DE	BLEU↑ DE-EN
NAT-FT	10	18.66	22.42
NAT-FT	100	19.17	23.20
LT	10	22.50	/
ENAT	9	24.28	26.10
NAT-REG	9	**24.61**	28.90
imitate-NAT	9	24.15	27.28
Flowseq	30	23.48	28.40
Transformer (ours)	-	27.30	31.33
PNAT (ours)	9	24.48	**29.16**

- 达到或超越了工作完成时的一系列方法

位置正确性v.s.翻译质量

- HSP相当于Oracle的位置信息（+30 BLEU）
- 利用AR预测器可以得到提升位置信息的作用（+11 BLEU）
- 利用NAR预测器速度更快，但位置不够准确（+4 BLEU）

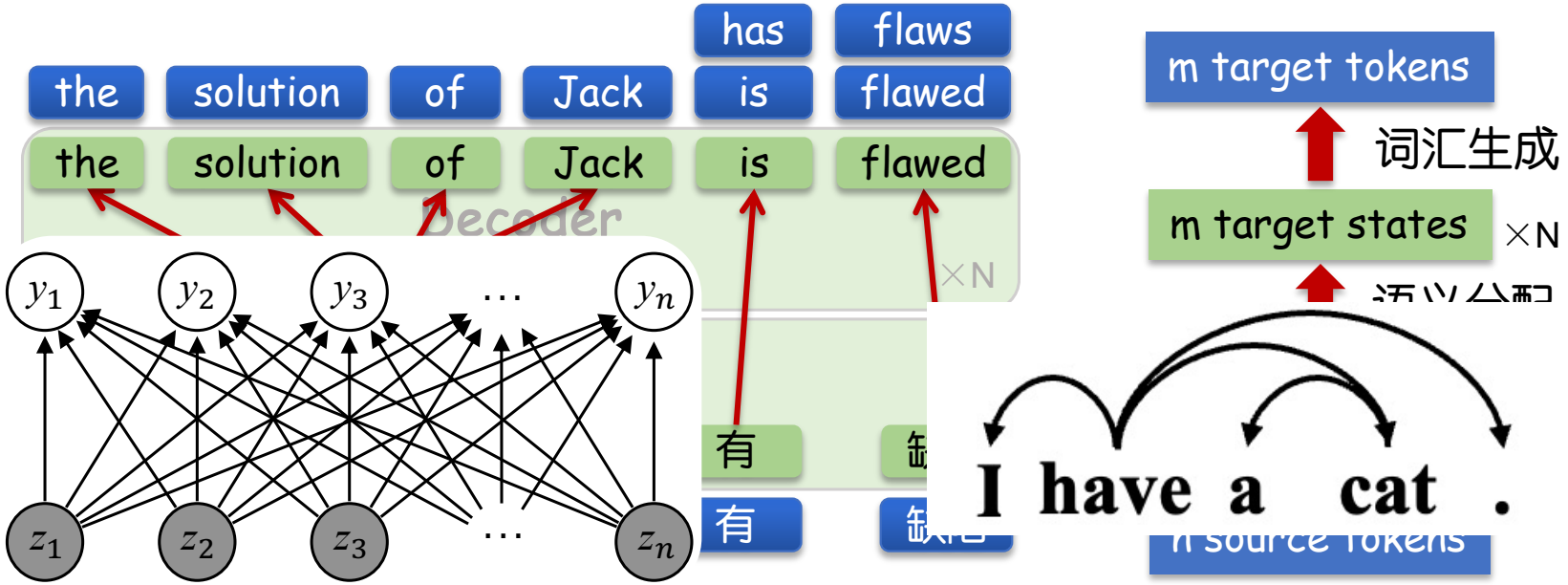
Model	Position Accuracy(%) absolute↑	relative↑	BLEU↑	Speed Up↑
Transformer (Beam=4)	/	/	30.68	1.0 ×
NAT	/	/	16.71	13.5×
PNAT w/ HSP	100.00	100.00	46.03	12.5×
w/ AR-Predictor	25.30	59.27	27.11	7.3 ×
w/ NAR-Predictor	23.11	55.57	20.81	11.7×

南京大學
NANJING UNIVERSITY

引入依存句法结构建模目标端的关联关系
INTERRELATION-BASED-NAT

文本之间的协同关系

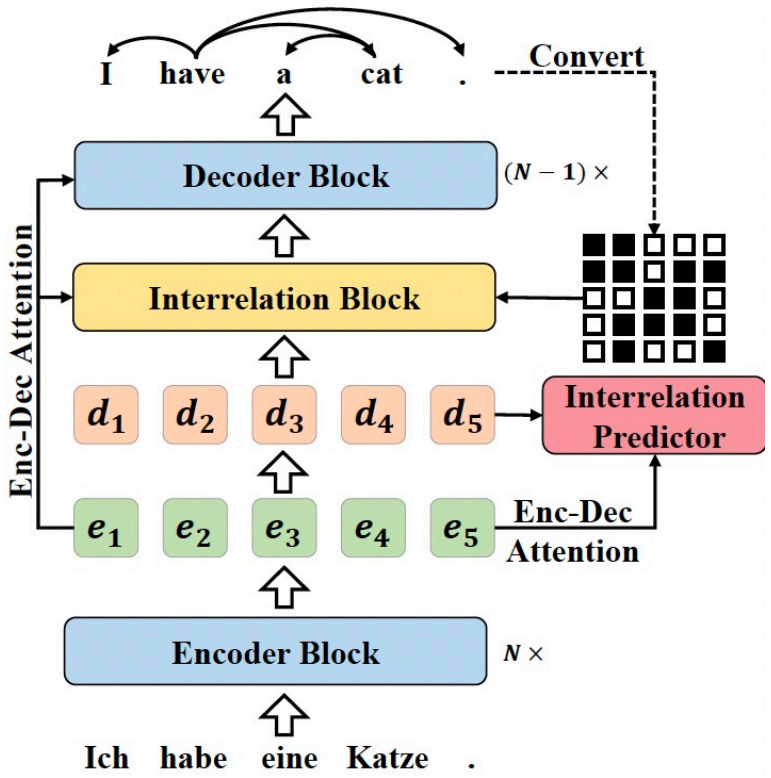
- 语义分配：决定target states（顺序、覆盖）
- 词汇生成：决定target token（一致、关联）



利用依存关系进行词汇生成过程中的协同！

我们的方案（inter-NAT）

- **基于依存关系提取不同解码位置之间的协同关系**
  - 描述了部分词语之间的依赖关系
- **基于协同关系进行后续并行解码**



基础实验结果

- 与基础NAT等模型的比较（without KD）

Model	WMT14		WMT16	
	EN-DE	DE-EN	EN-RO	RO-EN
SynST	20.74	25.50	/	/
Flowseq	20.85	25.40	29.86	30.69
AXE	20.40	24.90	30.47	31.42
NAT †	11.60	16.15	21.40	19.94
Inter-NAT †	**21.78**	**27.40**	**30.79**	**31.47**

- 极大的提升了NAT的能力

综合实验

- 加入KD（left）及reranking（right）等技术

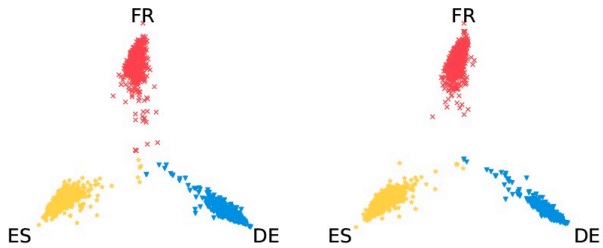
Model	WMT14 EN-DE	WMT14 DE-EN	WMT16 EN-RO	WMT16 RO-EN
NAT-FT	17.69	21.47	27.29	29.06
LT	19.80	/	/	/
NAT-REG	20.65	24.77	/	/
NAT-DCRF	23.44	27.22	/	/
ENAT	20.65	23.03	30.08	/
imitate-NAT	22.44	25.67	28.61	28.90
Flowseq	23.72	28.39	29.73	30.72
GLAT	25.21	29.84	31.19	**32.04**
Transformer †	27.25	31.53	33.97	33.60
Inter-NAT †	**25.78**	**29.89**	**32.10**	30.97

Model	N	WMT14 EN-DE	WMT14 DE-EN	WMT16 EN-RO	WMT16 RO-EN
NAT-FT	100	19.17	23.20	29.79	31.44
LT	10	21.00	/	/	/
NAT-REG	9	24.61	28.90	/	/
NAT-DCRF	19	26.80	30.04	/	/
ENAT	9	24.28	26.10	34.51	/
Flowseq	30	25.31	30.68	32.20	32.84
imitate-NAT	7	24.15	27.28	31.45	31.81
GLAT	7	26.55	31.02	32.87	33.51
Fully-NAT	/	**27.20**	31.39	33.71	**34.16**
Transformer †	-	27.25	31.53	33.97	33.60
Inter-NAT †	7	27.17	**31.45**	**33.75**	32.72

- 达到或超越了工作完成时的一系列方法

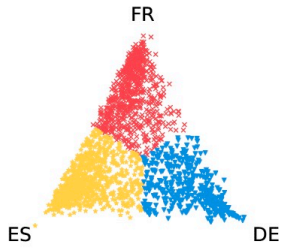
减少模态的能力

- **Inter-NAT能够增强词汇生成过程中进行协同的能力**
  - 减少词汇分配歧义（multi-modes）的场景 [Zhou et al. 2018]

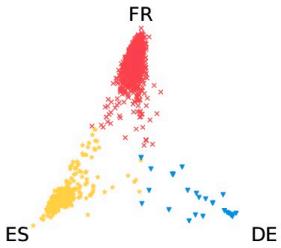


(a) Real data.

(b) Transformer



(c) NAT



(d) Inter-NAT

Data	$C(d)$	Euclidean distance
Real Data	0.78	0.22
Transformer †	0.56	0.23
NAT †	0.62	0.46
Inter-NAT †	0.41	0.27

总结

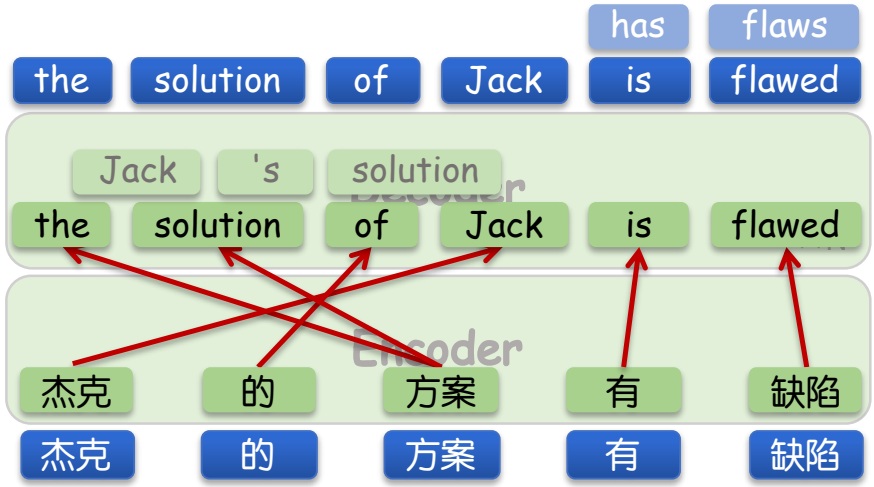
生成过程中，文本内部存在怎样的相互关系?

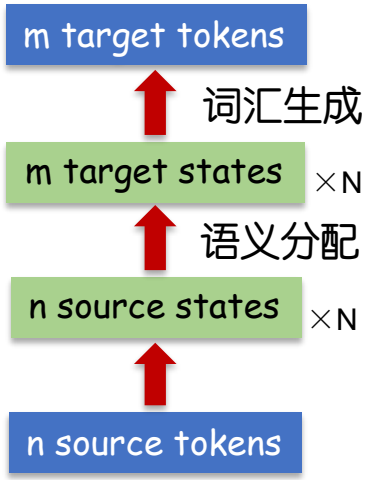
离散少量隐变量

调序关系

依赖关系

……





参考文献

- Warren Weaver. Translation. 1949
- Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers. 1984
- P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematic of statistical machine translation: Parameter estimation. Computational Linguistics 1993
- Sutskever Ilya, Vinyals Oriol, Le Quoc V.. Sequence to Sequence Learning with Neural Networks. NIPS 2014
- Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. EMNLP 2014

参考文献

- Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015
- Thang Luong, Kyunghyun Cho, Christopher Manning. Neural Machine Translation. ACL tutorial. 2016
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. NIPS 2017.
- Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation. arXiv preprint arXiv:1711.02281, 2017.
- Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. EMNLP 2018.
- Chunqi Wang, Ji Zhang, Haiqing Chen. Syntactically Supervised Transformers for Faster Neural Machine Translation. arXiv:1906.02780. 2019

参考文献

- Lukasz Kaiser, Samy Bengio, Aurko Roy, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, and Noam Shazeer. Fast decoding in sequence models using discrete latent variables. ICML 2018.
- Nader Akoury, Kalpesh Krishna, and Mohit Iyyer. Syntactically supervised transformers for faster neural machine translation. ACL 2019.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML 2001.
- Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, and Xu Sun. Imitation learning for non-autoregressive neural machine translation. ACL 2019.
- Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li.
- Glancing transformer for non-autoregressive neural machine translation. arXiv preprint arXiv:2008.07905. 2020.

参考文献

- Dozat T, Manning C D. Deep biaffine attention for neural dependency parsing. arXiv preprint arXiv:1611.01734, 2016.
- Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations. arXiv preprint arXiv:1803.02155, 2018.
- Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In 8th Inter-national Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.