Tencent AI Lab

Document-Level Neural Machine Translation 2

Longyue Wang

Natural Language Processing Centre

AI Lab

Tencent AI Lab

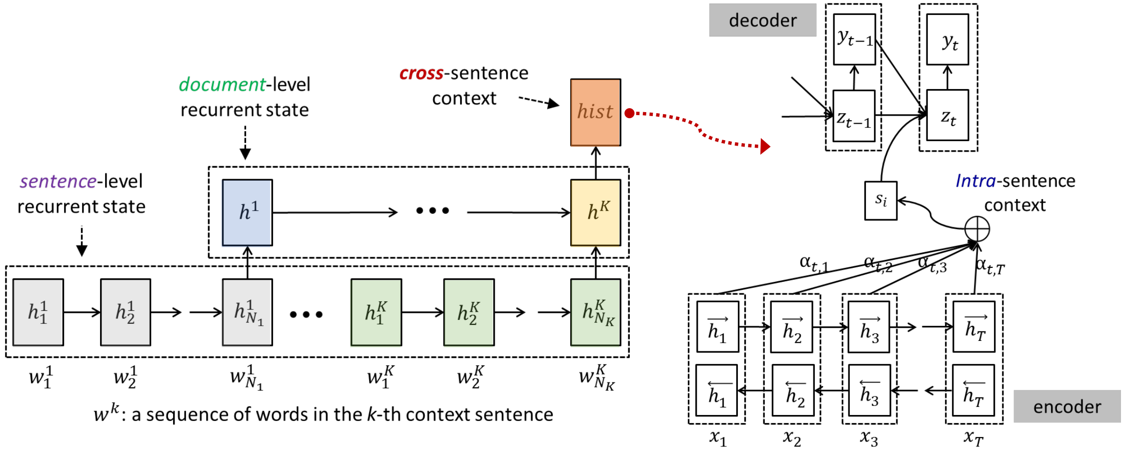
Agenda

- **Review**
- **Background**
  - Machine Translation
  - Discourse
- **Introduction**
- **Cross-Sentence Neural Machine Translation Models**
- **Comparison to Related Work**
  - Multi-Attention
  - Cache Memory
- **Document-Level MT Evaluation**
- **Conclusion**

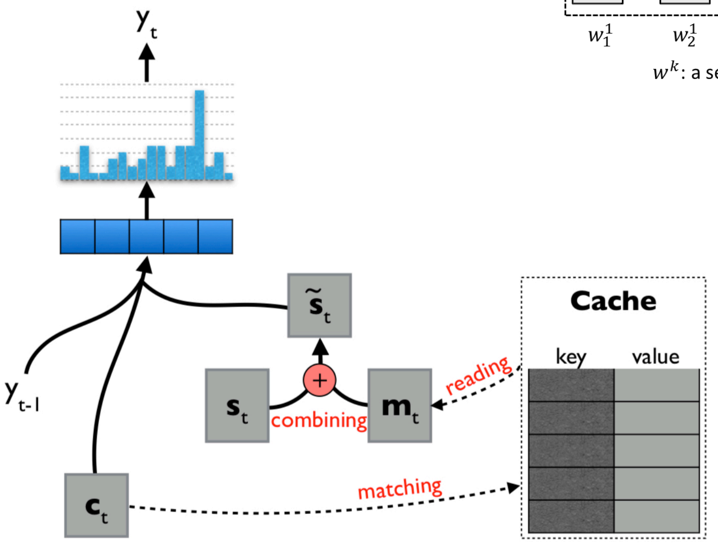
Tencent AI Lab

Review

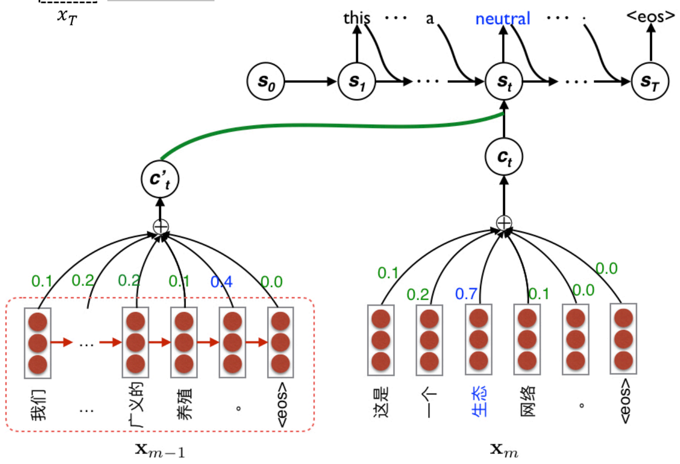
DNMT in Last Paper Reading



Wang et al. (2017)



Tu et al. (2017)



Jean et al. (2017)

Comparison

Work	Lang.	Len.	Prev. Rep.	Models	Integration
Jean et al. 2017	src	1	No	RNN	Att aux.
Wang et al. 2017	src	3	No	RNN	Init, Aux, Gate
Tu et al. 2017	src + trg	3+	Yes	RNN	Aux,
Miculicich et al. 2018	src + trg	3	yes	Transformer	Att aux., multi-head
Zhang et al., 2018	src	3+	No	Transformer	Enc, Dec, Gate
Voita et al., 2018	--	--	--	Transformer	--

Tencent
AI Lab

Hierarchical Attention Networks

Tencent AI Lab

Overall

Document-Level Neural Machine Translation with Hierarchical Attention Networks

Lesly Miculicich[†‡] Dhananjay Ram[†‡] Nikolaos Pappas[†] James Henderson[†]
[†] Idiap Research Institute, Switzerland
[‡] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{lmiculicich, dram, npappas, jhenderson}@idiap.ch

Main Contributions:

- HAN framework to capture cross-sentence context in a structured and **dynamic** manner.
- **Transformer** and strong baselines.
- Ablation study to show **source** and **target** sides are complementary.

Approach

HAN has **two levels** of abstraction:

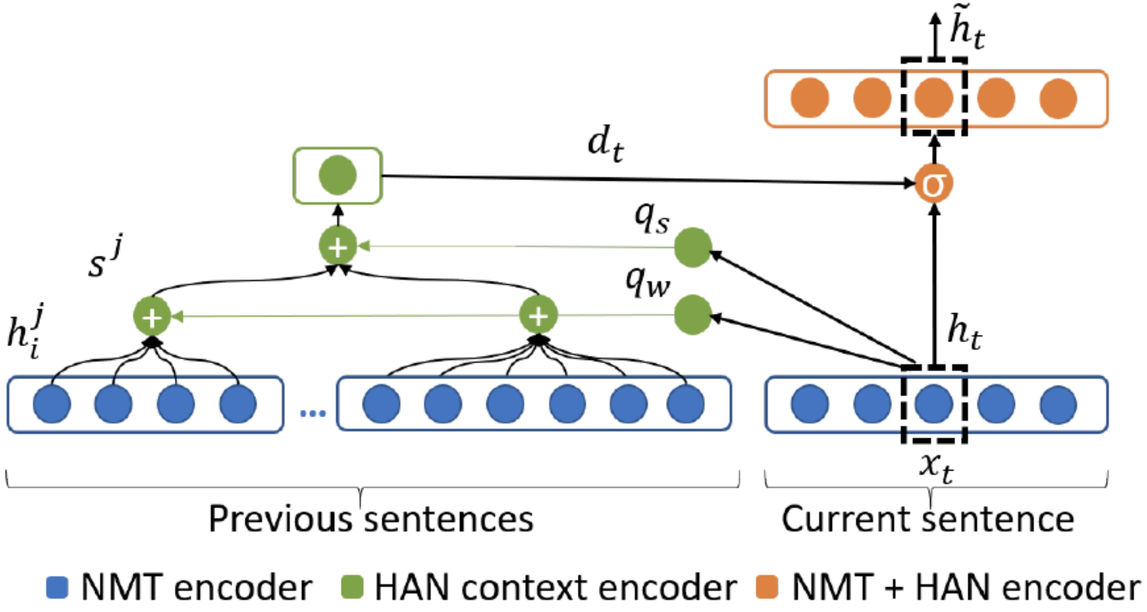
- **Word-Level** abstraction summarizes information from each previous sentence j into a vector s^j
- **Sentence-Level** abstraction summarizes the contextual information required at time t in d_t
- **Context Gating** regulates the information at h_t and d_t

$$q_w = f_w(h_t)$$

$$s^j = \text{MultiHead}(q_w, h_i^j)$$

$$q_s = f_s(h_t)$$

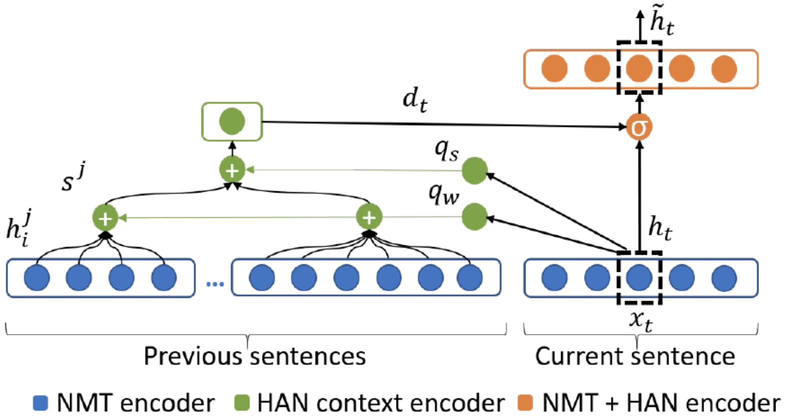
$$d_t = \text{FFN}(\text{MultiHead}(q_s, s^j))$$



Approach

Integrate the **global context** into NMT with **five strategies**:

- **HAN encoder**: query (h_{xt}), value (previous encoded states h_{xi}^{j}) at encoding time
- **HAN decoder source**: : query (h_{yt}), value (previous encoded states h_{xi}^{j}) at decoding time
- **HAN decoder**: : query (h_{yt}), value (previous decoded states h_{yi}^{j}) at decoding time
- **HAN decoder alignment**: : query (h_{yt}), value (alignment vector c_i^{j}) at decoding time
- **Combination**: combine source-target by joining HAN encoder and HAN decoder



Experiments

Training corpus consists of **three domains** and **two language** pairs:

- TED, Subtitle and News
- Chinese-English and Spanish-English

	TED Talks		**Subtitles**		**News**
	Zh–En	Es–En	Zh–En	Es–En	Es–En
Training	0.2M	0.2M	2.2M	4.0M	0.2M
Development	0.8K	0.8K	1.1K	1.0K	1.9K
Test	5.5K	4.7K	1.2K	1.0K	13.5K

Configurations:

- base model
- K = 3

Results

Translation performance:

- **Transformer** is better than previously published models?
- All proposed models perform at least as well as **cache-base NMT**
- The best scores are obtained by **combining** encoder and decoder HAN models
- HAN **encoder** is second best and **decoder** may contain erroneous predictions

	TED Talks				Subtitles				News	
	Zh–En		Es–En		Zh–En [4]		Es–En		Es–En	
Models	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ	BLEU	Δ
NMT transformer	16.87		35.44		28.60		35.20		21.36	
+ cache (Tu et al., 2018)	17.32	$(+0.45)^{***}$	36.46	$(+1.02)^{***}$	28.86	$(+0.26)$	35.49	$(+0.29)$	22.36	$(+1.00)^{***}$
+ HAN encoder	17.61	$(+0.74)^{***}_{\dagger\dagger}$	36.91	$(+1.47)^{***}_{\dagger\dagger}$	29.35	$(+0.75)^{*}_{\dagger}$	35.96	$(+0.76)^{*}_{\dagger}$	22.36	$(+1.00)^{***}$
+ HAN decoder	17.39	$(+0.52)^{***}$	37.01	$(+1.57)^{***}_{\dagger\dagger\dagger}$	29.21	$(+0.61)^{*}$	35.50	$(+0.30)$	22.62	$(+1.26)^{***}_{\dagger\dagger\dagger}$
+ HAN decoder *source*	17.56	$(+0.69)^{***}_{\dagger\dagger}$	36.94	$(+1.50)^{***}_{\dagger\dagger}$	28.92	$(+0.32)$	35.71	$(+0.51)^{*}$	22.68	$(+1.32)^{***}_{\dagger\dagger\dagger}$
+ HAN decoder *alignment*	17.48	$(+0.61)^{***}_{\dagger}$	37.03	$(+1.60)^{***}_{\dagger\dagger\dagger}$	28.87	$(+0.27)$	35.63	$(+0.43)$	22.59	$(+1.23)^{***}_{\dagger\dagger\dagger}$
+ HAN encoder + HAN decoder	**17.79**	$(+0.92)^{***}_{\dagger\dagger\dagger}$	**37.24**	$(+1.80)^{***}_{\dagger\dagger\dagger}$	**29.67**	$(+1.07)^{**}_{\dagger}$	**36.23**	$(+1.03)^{**}_{\dagger\dagger}$	**22.76**	$(+1.40)^{***}_{\dagger\dagger\dagger}$

Analysis

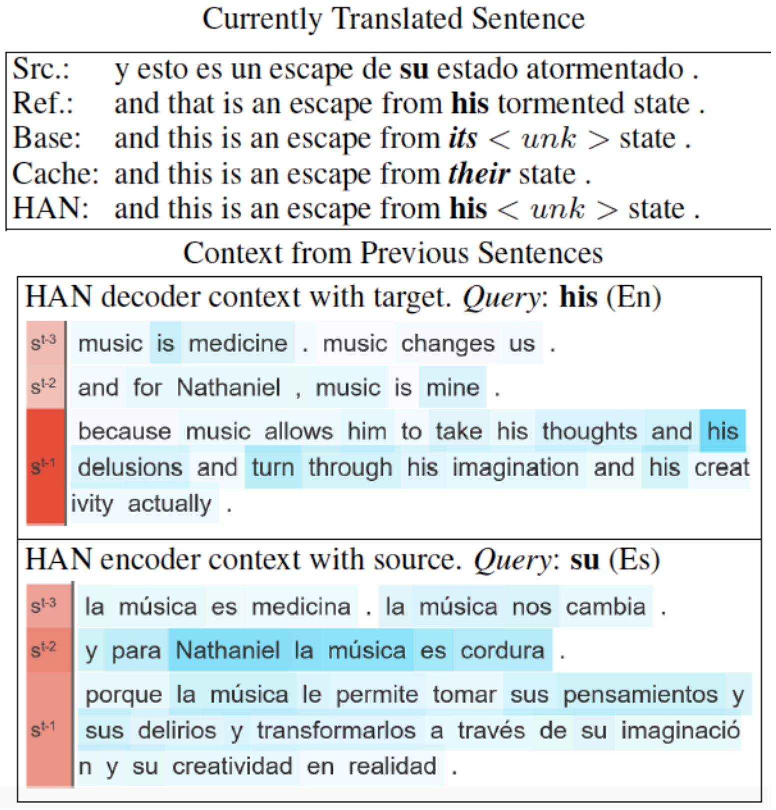
Effect of K:

- Best performance for TED talks and news is archived with 3
- Subtitles needs more history 3 and 7.

Visualization:

- HAN correctly translates the ambiguous Spanish pronoun “su” into the English “his”.
- HAN decoder highlighted a previous mention of “his”, and the HAN encoder highlighted the antecedent “Nathaniel”.

	TED Talks		**Subtitles**		**News**
k	Zh–En	Es–En	Zh–En	Es–En	Es–En
1	17.70	37.20	29.35	36.20	22.46
3	**17.79**	**37.24**	29.67	**36.23**	**22.76**
5	17.49	37.11	**29.69**	36.22	22.54
7	17.00	37.22	29.64	36.21	22.64



Analysis

Accuracy of Pronoun/Noun Translation:

- For **nouns**, the joint HAN achieves the best accuracy.
- For **pronouns**, the joint model has the best result for TED talks and news.
- For **subtitles**, HAN encoder is better and HAN decoder produces mistakes.
- Subtitles is a challenging corpus for personal pronoun disambiguation - multiple speakers.

	Noun Translation					**Pronoun Translation**				
	TED Talks		**Subtitles**		**News**	**TED Talks**		**Subtitles**		**News**
Model	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En
NMT Transformer	40.16	65.97	46.65	61.79	47.94	63.44	68.00	69.71	65.83	47.22
+ cache	40.87	66.75	46.00	61.87	49.91	63.53	68.66	69.97	66.27	49.34
+ HAN encoder	41.93	67.75	46.78	61.52	50.06	64.05	69.17	**71.04**	**68.56**	49.57
+ HAN decoder	41.61	67.35	46.78	61.99	50.03	64.02	69.36	70.50	67.03	49.33
+ HAN encoder + HAN decoder	**42.99**	**67.81**	**47.43**	**62.30**	**50.40**	**64.35**	**69.60**	70.60	67.47	**49.59**

Tencent AI Lab

Analysis

Cohesion and Coherence Evaluation:

- **Lexical cohesion**: Wong and Kit (2012).
- HAN decoder achieves the best score because it produces a larger quantity of repetitions.
- The scores are still far from the human reference.
- **Coherence**: Latent Semantic Analysis (LSA) (Foltz et al., 1998).
- Joint HAN model consistently obtains the best coherence score, but close to other HAN models.

	Lexical cohesion					Coherence				
NMT Transformer	54.26	51.98	51.87	51.77	30.06	0.298	0.299	0.283	0.262	0.279
+ HAN encoder	54.87	52.35	51.89	52.33	30.34	0.304	0.299	0.285	0.262	0.280
+ HAN decoder	54.95	**52.43**	**52.33**	52.43	30.41	0.302	0.301	0.287	0.265	0.282
+ HAN enc. + HAN dec.	**55.40**	52.36	51.94	**52.75**	**30.58**	**0.305**	**0.302**	**0.287**	**0.265**	**0.282**
Human reference	56.08	57.02	54.81	58.19	35.12	0.310	0.314	0.296	0.270	0.298

Document-Level Transformer

Overall

Improving the Transformer Translation Model
with Document-Level Context

Jiacheng Zhang[†], Huanbo Luan[†], Maosong Sun[†], FeiFei Zhai[#],
Jingfang Xu[#], Min Zhang[§] and Yang Liu[†‡*]
[†]Institute for Artificial Intelligence
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[‡]Beijing National Research Center for Information Science and Technology
[#]Sogou Inc., Beijing, China
[§]Soochow University, Suzhou, China

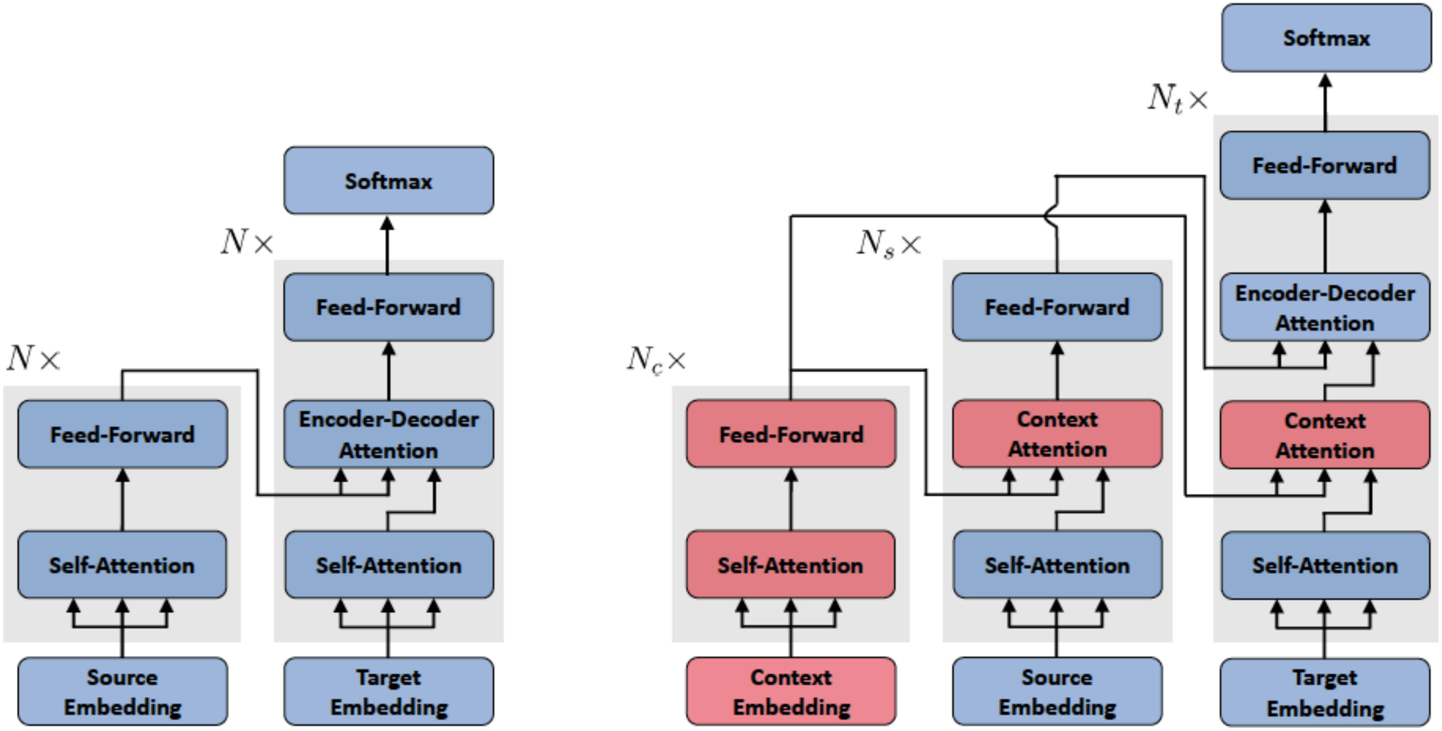
Main Contribution:

- Extend the **Transformer** with a new context encoder to represent document-level context.
- Introduce a **two-step training** method to take full advantage of abundant sentence-level parallel corpora and limited document-level parallel corpora.

Approach

Architecture:

- Use **multi-head self-attention** to compute the representation of document-level context.
- X_c is the **concatenation** of all vector representations of all source contextual words.
- $A(1) = \mathrm{MultiHead}(Xc;Xc;Xc)$; $Q = K = V = X_c$



Approach

Document-level Context Integration:

- Integration into the Encoder
- Integration into the Decoder
- Context Gating

$$\mathbf{D}^{(n)} = \text{MultiHead}\left(\mathbf{B}^{(n)}, \mathbf{C}^{(N_c)}, \mathbf{C}^{(N_c)}\right)$$

$$\mathbf{F}^{(n)} = \text{MultiHead}\left(\mathbf{E}^{(n)}, \mathbf{C}^{(N_c)}, \mathbf{C}^{(N_c)}\right)$$

$$\text{Gating}(\mathbf{H}) = \lambda\mathbf{H} + (1 - \lambda)\text{SubLayer}(\mathbf{H})$$

$$\lambda = \sigma(\mathbf{W}_i\mathbf{H} + \mathbf{W}_s\text{SubLayer}(\mathbf{H}))$$

Approach

Pre-training:

- In the **first** step, sentence-level parameters θs are estimated on the combined sentence-level parallel corpus, but newly introduced modules are inactivated:

$$\hat{\boldsymbol{\theta}}_s = \operatorname*{argmax}_{\boldsymbol{\theta}_s} \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in D_s \cup D_d} \log P(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}_s)$$

- In the **second** step, document-level parameters θd are estimated on the document-level parallel corpus D_d

$$\hat{\boldsymbol{\theta}}_d = \operatorname*{argmax}_{\boldsymbol{\theta}_d} \sum_{\langle \mathbf{X}, \mathbf{Y} \rangle \in D_d} \log P(\mathbf{Y} | \mathbf{X}; \hat{\boldsymbol{\theta}}_s, \boldsymbol{\theta}_d)$$

- Our approach keeps θs fixed when estimating θd

Experiments

Training corpus consists of three domains and two language pairs:

- Chinese-English: 41K documents, 2M sentence pairs
- French-English: 1,824 documents with 220K sentence pairs

Configurations:

- base model
- K= 3+
- Thumt

Results

Effect of length:

- using **2 preceding source** sentences achieves the best translation performance.
- Using **more** preceding sentences does not bring any improvement and increases computational cost.

Effect of layer:

- using only **one SAN layer** suffices to achieve good performance

# sent.	1	2	3
MT06	49.38	**49.69**	49.49

Table 1: Effect of context length on translation quality. The BLEU scores are calculated on the development set.

# Layer	MT06
1	**49.69**
2	49.38
3	49.54
4	49.59
5	49.31
6	49.43

Table 2: Effect of self-attention layer number (i.e., N_c) on translation quality. The BLEU scores are calculated on the development set.

Tencent AI Lab

Results

Main results:

- The approach achieves significant improvements **over the original Transformer** model

Method	Model	MT06	MT02	MT03	MT04	MT05	MT08	All
(Wang et al., 2017)	RNNsearch	37.76	-	-	-	36.89	27.57	-
(Kuang et al., 2017)	RNNsearch	-	34.41	-	38.40	32.90	31.86	-
(Vaswani et al., 2017)	Transformer	48.09	48.63	47.54	47.79	48.34	38.31	45.97
(Kuang et al., 2017)*	Transformer	48.14	48.97	48.05	47.91	48.53	38.38	46.37
this work	Transformer	**49.69**	**50.96**	**50.21**	**49.73**	**49.46**	**39.69**	**47.93**

Method	Dev	Test
Transformer	29.42	35.15
this work	**30.40**	**36.04**

Analysis

Subjective Evaluation:

- Human evaluation on 198 sentences

	>	=	<
Human 1	24%	45%	31%
Human 2	20%	55%	25%
Human 3	12%	52%	36%
Overall	19%	51%	31%

Effect of **Two-Step Training**

- The **fourth and fifth rows** use the two-step strategy to take advantage of both sentence- and document-level parallel corpora

sent.	doc.	MT06	MT02	MT03	MT04	MT05	MT08	All
940K	-	36.20	42.41	43.12	41.02	40.93	31.49	39.53
2M	-	48.09	48.63	47.54	47.79	48.34	38.31	45.97
-	940K	34.00	38.83	40.51	38.30	36.69	29.38	36.52
940K	940K	37.12	43.29	43.70	41.42	41.84	32.36	40.22
2M	940K	49.69	50.96	50.21	49.73	49.46	39.69	47.93

Analysis

Effect of Context Integration:

Integration	MT06	MT02	MT03	MT04	MT05	MT08	All
none	48.09	48.63	47.54	47.79	48.34	38.31	45.97
encoder	48.88	50.30	49.34	48.81	49.75	39.55	47.51
decoder	49.10	50.31	49.83	49.35	49.29	39.07	47.48
both	49.69	50.96	50.21	49.73	49.46	39.69	47.93

Effect of Context Gating:

Gating	MT06	MT02	MT03	MT04	MT05	MT08	All
w/o	49.33	50.56	49.74	49.29	50.11	39.02	47.55
w/	49.69	50.96	50.21	49.73	49.46	39.69	47.93

Tencent AI Lab

Thank You

Longyue Wang
NLP Centre
Tencent AI Lab
vincentwang0229@gmail.com