Security Level: Public

Document-level Machine Translation: the Current State and the Challenges

Qun Liu, Liangyou Li, Mingzhou Xu
Huawei Noah's Ark Lab

www.huawei.com

An Invited Talk at DiscoMT 2019 Workshop
3 November 2019, Hong Kong

HUAWEI TECHNOLOGIES CO., LTD.

Preface

- Huge progresses have been made in MT by applying DP.
- Some paper even claimed to achieve Human Parity.
- Is MT a solved problem? No!!!
  - Document Level MT
  - Domain MT
  - Low-resource MT
  - Multi-lingual MT
  - Trustworthy MT
  - ………

Content

Sentence-Level vs Document-Level

- Samuel Läubli; **Rico Sennrich**; Martin Volk (2018). [Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation.]() In Proceedings of the EMNLP2018 pp. 4791-4796.

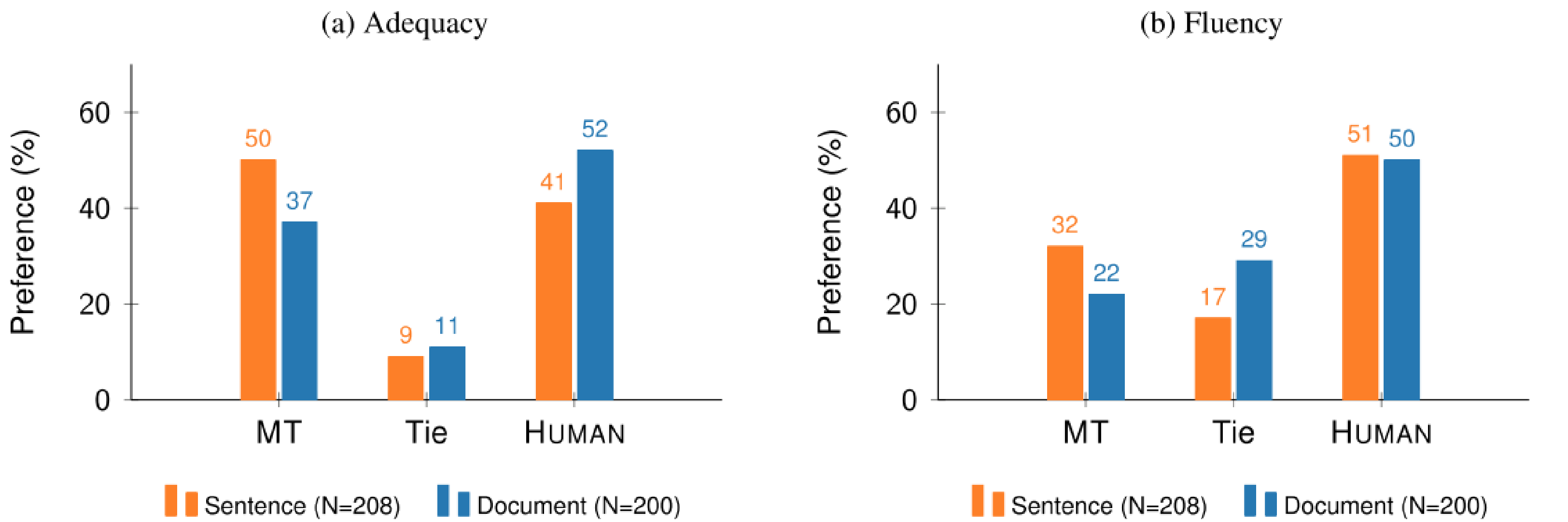


Figure 1: Raters prefer human translation more strongly in entire documents. When evaluating isolated sentences in terms of adequacy, there is no statistically significant difference between HUMAN and MT; in all other settings, raters show a statistically significant preference for HUMAN.

Source: 微信挪车
HUMAN: WeChat Move the Car
MT: Twitter Move Car / WeChat mobile / WeChat Move

Example #1

有一个寒冷的冬夜，农夫在路边拾到一条被冻僵的蛇，他觉得蛇很可怜，于是便把蛇搂在怀中，为它取暖。当蛇醒来，竟向着农夫的胸口大力一咬，令他中毒死亡。

On a cold winter night, the farmer picked up a frozen snake on the side of the road. He felt that the snake was very poor, so he put the snake in his arms and warmed it. When the snake woke up, he bit a bit to the farmer's chest and poisoned him.

Article Errors

Example #2 – Chinese Source

HUAWEI

“舒克，你都大了，可以自己出去找东西吃了。”一天，妈妈对小老鼠舒克说。 “真的吗？”舒克高兴了。 舒克是一只生活在中国的小老鼠，他从生下来以后就一直憋在洞里，从来没有出去玩过。 “今大晚上，我带你出去，先认认路，以后你就可以自己去了。”妈妈一边说，一边磨牙。 舒克也学着妈妈的样子磨牙。他爱吃好东西。每次妈妈给他带回来好吃的．他都吃个没够。 夜里，舒克跟在妈妈身后出了洞。

Example #2 – Human Translation

"Shuk, you are grown up. You can go out to find food by yourself." Shuk, a small mouse, was told by his mother one day. "Really?" Shuk felt delighted. Shuk was a small mouse living in China. He had been hold within his hole since he was born and had never gone out of the hole. "Tonight, I will take you out to recognize the road first, then you can go out by yourself." The mother said while grinding her teeth. Shuk also ground his teeth, imitating his mother. He loved to eat yummy food. Every time his mother brought taste things back, he always enjoyed the food and felt unsatisfied. At night, Shuk went out of the hole, following his mother.

Example #2 – English MT by Google

"Shuke, you are all big, you can go out and find something to eat yourself." One day, my mother said to the little mouse Shuk. "Really?" Shuk was happy. Shuk is a small mouse living in China. He has been lying in a hole since he was born and never went out to play. "Tonight, I will take you out, first recognize the road, and then you can go by yourself." Mother said while grinding his teeth. Shuk also learned how to make a mother's teeth. He loves to eat good things. Every time my mother brought him back to eat delicious. He didn't have enough to eat. At night, Shuk had a hole behind his mother.

Example #2 – English MT by Google

"Shuke, you are all big, you can go out and find something to eat yourself." One day, my mother said to the little mouse Shuk. "Really?" Shuk was happy. Shuk is a small mouse living in China. He has been lying in a hole since he was born and never went out to play. "Tonight, I will take you out, first recognize the road, and then you can go by yourself." Mother said while grinding his teeth. Shuk also learned how to make a mother's teeth. He loves to eat good things. Every time my mother brought him back to eat delicious. He didn't have enough to eat. At night, Shuk had a hole behind his mother.

Inconsistent Proper Noun Translation

Example #2 – English MT by Google

"Shuke, you are all big, you can go out and find something to eat yourself." One day, my mother said to the little mouse Shuk. "Really?" Shuk was happy. Shuk is a small mouse living in China. He has been lying in a hole since he was born and never went out to play. "Tonight, I will take you out, first recognize the road, and then you can go by yourself." Mother said while grinding his teeth. Shuk also learned how to make a mother's teeth. He loves to eat good things. Every time my mother brought him back to eat delicious. He didn't have enough to eat. At night, Shuk had a hole behind his mother.

Tense Errors

Example #2 – English MT by Google

"Shuke, you are all big, you can go out and find something to eat yourself." One day, my mother said to the little mouse Shuk. "Really?" Shuk was happy. Shuk is a small mouse living in China. He has been lying in a hole since he was born and never went out to play. "Tonight, I will take you out, first recognize the road, and then you can go by yourself." Mother said while grinding his teeth. Shuk also learned how to make a mother's teeth. He loves to eat good things. Every time my mother brought him back to eat delicious. He didn't have enough to eat. At night, Shuk had a hole behind his mother.

Dropped Pronoun Translation Errors

Example #2 – English MT by Google

"Shuke, you are all big, you can go out and find something to eat yourself." One day, my mother said to the little mouse Shuk. "Really?" Shuk was happy. Shuk is a small mouse living in China. He has been lying in a hole since he was born and never went out to play. "Tonight, I will take you out, first recognize the road, and then you can go by yourself." Mother said while grinding his teeth. Shuk also learned how to make a mother's teeth. He loves to eat good things. Every time my mother brought him back to eat delicious. He didn't have enough to eat. At night, Shuk had a hole behind his mother.

Inconsistent Verb Phrases

Errors in Document-Level MT

- Rachel Bawden, Rico Sennrich, Alexandra Birch, Barry Haddow. *Evaluating Discourse Phenomena in Neural Machine Translation.* **NAACL** 2018

Errors in Document-Level MT

Source:

context:	Oh, I hate **flies**. Look, there's another one!
current sent.:	Don't worry, I'll kill **it** for you.

Target:

1	context:	Ô je déteste les **mouches**. Regarde, il y en a une autre !
	correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
2	context:	Ô je déteste les **moucherons**. Regarde, il y en a un autre !
	correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.
3	context:	Ô je déteste les **araignées**. Regarde, il y en a une autre !
	semi-correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
4	context:	Ô je déteste les **papillons**. Regarde, il y en a un autre !
	semi-correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.

Figure 1: Example block from the coreference set.

Errors in Document-Level MT

Source:

context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

Target:

context:	Qu'est-ce qu'il y a de **dingue** chez moi ?
correct:	Est-ce que ça c'est **dingue** ?
incorrect:	Est-ce que ça c'est fou ?

Source:

context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

Target:

context:	Qu'est-ce qu'il y a de **fou** chez moi ?
correct:	Est-ce que ça c'est **fou** ?
incorrect:	Est-ce que ça c'est dingue ?

Figure 2: Example block from the coherence/cohesion test: alignment.

Errors in Document-Level MT

Source:

context:	So what do you say to £50?
current sent.:	It's a little **steeper** than I was expecting.

Target:

context:	Qu'est-ce que vous en pensez de 50£ ?
correct:	C'est un peu plus **cher** que ce que je pensais.
incorrect:	C'est un peu plus **raide** que ce que je pensais.

Source:

context:	How are your feet holding up?
current sent.:	It's a little **steeper** than I was expecting.

Target:

context:	Comment vont tes pieds ?
correct:	C'est un peu plus **raide** que ce que je pensais.
incorrect:	C'est un peu plus **cher** que ce que je pensais.

Figure 3: Example block from the coherence/cohesion test: lexical disambiguation.

Errors in Document-Level MT

- Elena Voita, Rico Sennrich, Ivan Titov. *When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion*. **ACL** 2019

  - Deixis
  - Ellipsis
  - Lexical Cohesion

Document-Level Translation Errors

- Phenomena
  - Inconsistent errors (not consistent)
  - Incohesive errors (not integrated lexically)
  - Incoherent errors (not integrated structurally)
- In document level MT, the differences between these types of errors are subtle.

Document-Level Translation Errors

- Taxonomy for Document-Level Translation Errors
  – Errors Caused by Missing Elements which are sensitive to the context
    - Missing Pronouns (Dropped Pronouns, Ellipsis, Zero Anaphora)
    - Missing Noun Phrases
    - Missing Articles
    - Missing Tense
  – Errors Caused by Ambiguous Words which are sensitive to the context
    - Ambiguous Pronouns (Deixis, Anaphora)
    - Ambiguous Noun Phrases (Lexical Cohesion)
    - Ambiguous Verb Phrases (Lexical Cohesion)
    - Ambiguous Tense

Content

Approaches for Document Level MT

- Pre-processing Approaches
- Post-processing Approaches
- RNN-based Document-Level MT Models
- Transformer-based Document-Level MT Models

Pre-processing Approaches

- Longyue Wang, Zhaopeng Tu, Andy Way, Qun Liu. *Learning to Jointly Translate and Predict Dropped Pronouns with a Shared Reconstruction Mechanism*. **EMNLP** 2018.
- Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, Qun Liu. *Translating Pro-Drop Languages with Reconstruction Models*. **AAAI** 2018.
- Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Siyou Liu, Hang Li, Andy Way and Qun Liu. *A Novel and Robust Approach for Pro-Drop Language Translation*. **Machine Translation**. 31.1-2 (2017): 65-87.
- Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way and Qun Liu. *A Novel Approach for Dropped Pronoun Translation*. **NAACL-HLT** 2016.

Predict Dropped Pronouns by Word Alignment

- **Motivation:** To improve the translation for dropped pronouns, we try to restore the dropped pronouns in the source side.
- **Challenge:** Lack of annotated corpus for dropped pronouns restoration.
  - Chinese PennTreebank contains empty category annotations but its size is rather rare
- **Our idea**: large parallel corpara are available and can be used to provide hints of dropped pronouns by using word alignment.

1 (a) （DP） 喜欢 这份 工作 吗？

1 (b) Do **you** like this job ?

2 (a) 是的，（DP）很喜欢（DP），谢谢 （DP）。

2 (b) Yes , **I** like **it** . Thank **you** .

1 (a) （你） 喜欢 这份 工作 吗？

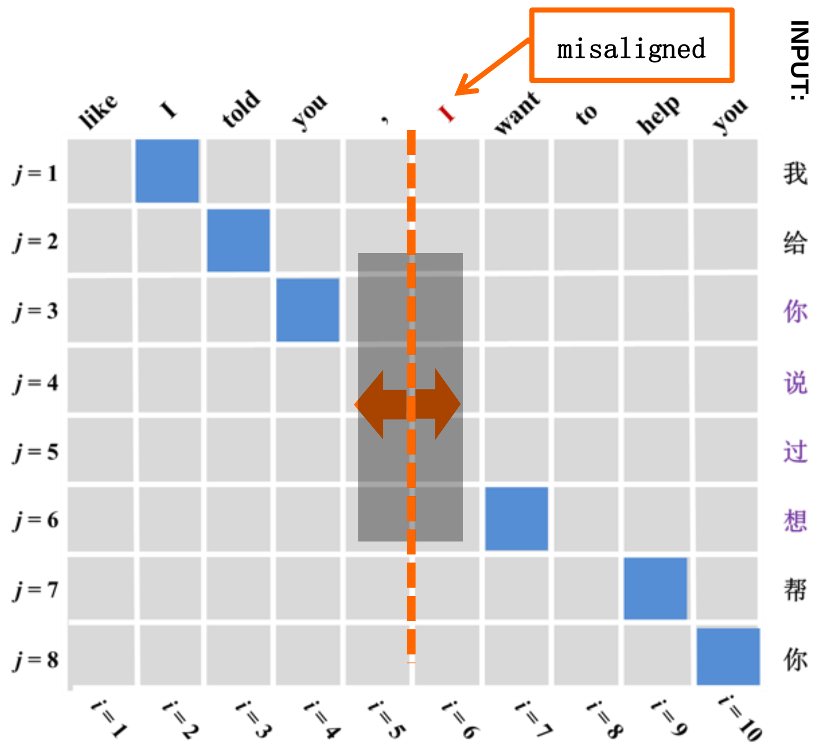
1 (b) Do **you** like this job ?

2 (a) 是的，（我）很喜欢（它），谢谢 （你）。

2 (b) Yes , **I** like **it** . Thank **you** .

Dropped Pronoun (DP) Training Corpus

- We automatically build a large DP training corpus:
  - Build word alignment for a large parallel corpus
  - Use bidirectional search algorithm to detect possible positions of DPs
  - To determine exact DP words, we use LM to score all possible sentences by inserting corresponding Chinese DPs

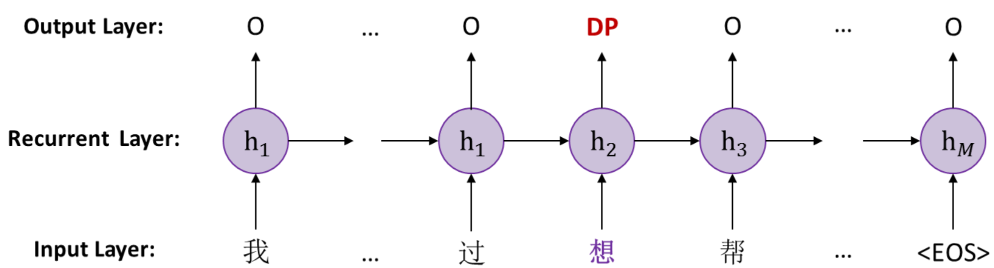


ID	possible positions to insert DP-I
1	我 给 你 DP-I 说 过 想 帮 你
2	我 给 你 说 DP-I 过 想 帮 你
3 ✔	我 给 你 说 过 DP-I 想 帮 你
4	我 给 你 说 过 想 帮 你

OUTPUT: 我 给 你 说 过 <DP>我</DP> 想 帮 你

DP Generation

- We train NN models on our DP training corpus, and split the task into two subtasks: DP Position detection (DPP) and DP Word prediction (DWP).
  - Regarding the DPP detection as a sequence labelling task, we employ RNNs to read Chinese sentence and output binary labels (DP/O)

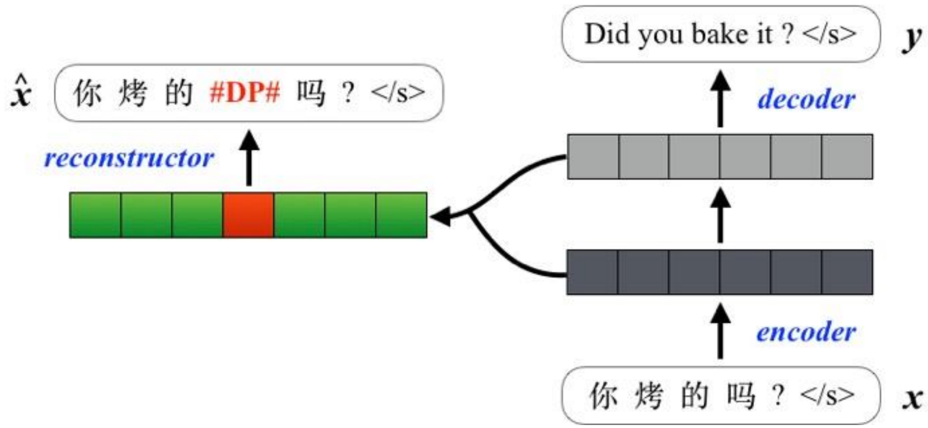


  - Based on DPPs, we then use MLP to predict DPW using rich features: lexical, syntax and inter-/intra-sentence context (Xiang et al., 2013; Yang et al., 2015)

ID.	Lexical Feature Set
1	W surrounding words around p
2	W surrounding POS tags around p
3	previous pronoun in the same sentence
4	following pronoun in the same sentence
	Context Feature Set
5	pronouns in previous X sentences
6	pronouns in following X sentences
7	Y nouns in previous sentences
8	Y nouns in following sentences
	Syntax Feature Set
9	path from current word (p) to the root
10	path from previous word ($p-1$) to the root

Shared Reconstructor

- The shared reconstructor reads from both the encoder and decoder hidden states, as well as the DP-annotated source sentence, and outputs a reconstruction score.



$$R(\hat{\mathbf{x}}|\mathbf{h}^{enc}, \mathbf{h}^{dec}) = \prod_{t=1}^{T} g_r(\hat{x}_{t-1}, \mathbf{h}_t^{rec}, \hat{\mathbf{c}}_t^{enc}, \hat{\mathbf{c}}_t^{dec})$$

$$\mathbf{h}_t^{rec} = f_r(\hat{x}_{t-1}, \mathbf{h}_{t-1}^{rec}, \hat{\mathbf{c}}_t^{enc}, \hat{\mathbf{c}}_t^{dec})$$

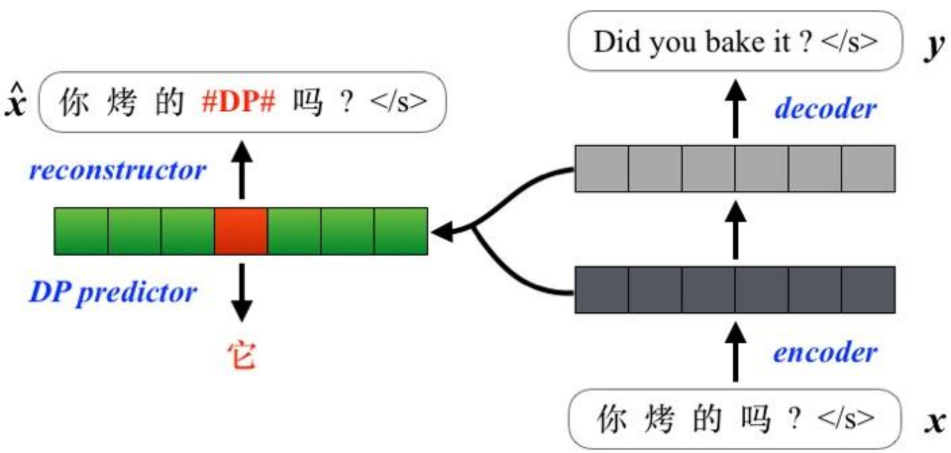
- For better interaction, we also propose the interactive attention feeds the context vector produced by one attention model to another attention model. e.g. enc→dec:

$$\hat{\alpha}^{enc} = \text{ATT}_{enc}(\hat{x}_{t-1}, \mathbf{h}_{t-1}^{rec}, \mathbf{h}^{enc})$$

$$\hat{\alpha}^{dec} = \text{ATT}_{dec}(\hat{x}_{t-1}, \mathbf{h}_{t-1}^{rec}, \mathbf{h}^{dec}, \hat{\mathbf{c}}_t^{enc})$$

Joint Prediction of Dropped Pronouns

- We leverage the DPPs predicted by an external model, which can achieve an accuracy of 88% in F1-score.
  - Transform the original DP prediction problem to DPW prediction given the pre-detected DPPs
  - Introduce an additional DPW-predition loss, which measures how well the DPW is generated from the corresponding hidden state in the reconstructor



$$J(\theta,\gamma,\psi) = \arg\max_{\theta,\gamma,\psi} \Big\{ \underbrace{\log L(\mathbf{y}|\mathbf{x};\theta)}_{likelihood} + \underbrace{\log R(\hat{\mathbf{x}}|\mathbf{h}^{enc},\mathbf{h}^{dec};\theta,\gamma)}_{reconstruction} + \underbrace{\log P(\mathbf{dp}|\hat{\mathbf{h}}^{rec};\theta,\gamma,\psi)}_{prediction} \Big\}$$

Training Objective

Experiments

- **Parallel Corpus**: 2M sentence pairs extracted from the US TV series subtitles.
- **DP Corpus** and **DPP Detector**: build them using the same approaches
- **Models**:
  - **Baseline**: standard NMT on original parallel corpus
  - **Baseline (+DPPs)**: a stronger baseline trained on the new parallel corpus (DPP-labelled source + target sentence pairs), which is evaluated on the DPP-labelled sentences (by DPP detector)
  - **Separate-Rec→(+DPs)**: best model in previous section
  - Our models:
    - **Shared-Rec (indep.)→(+DPPs)**: shared reconstructor
    - **+ joint**: shared reconstructor with DPW prediction
    - **enc→dec** or **dec→enc**: +interactive attention mechanism

Experiments

- **Baselines**: baseline trained on the DPP-annotated data outperforms the other two counterparts
- **Shared-Rec (indep.)→(+DPPs)**: shared reconstructo not only outperforms the baseline, but also surpasses its separate reconstructor counterpart
- **+ Joint**: introducing a joint prediction objective can achieve a further improvement of +0.61 BLEU
- + **Interactive Attention**: enc→dec interaction attention achieves the best performance

#	Model	#Params	Speed		BLEU
			Train	Decode	
	Existing system (Wang et al., 2018)				
1	Baseline	86.7M	1.60K	15.23	31.80
2	Baseline (+DPs)	86.7M	1.59K	15.20	32.67
3	Separate-Recs⇒(+DPs)	+73.8M	0.57K	12.00	35.08
	Our system				
4	Baseline (+DPPs)	86.7M	1.54K	15.19	33.18
5	Shared-Rec$_{independent}$⇒(+DPPs)	+86.6M	0.52K	11.87	35.27†‡
6	Shared-Rec$_{independent}$⇒(+DPPs) + joint prediction	+87.9M	0.51K	11.88	35.88†‡
7	Shared-Rec$_{enc\rightarrow dec}$⇒(+DPPs) + joint prediction	+91.9M	0.48K	11.84	**36.53**†‡
8	Shared-Rec$_{dec\rightarrow enc}$⇒(+DPPs) + joint prediction	+89.9M	0.49K	11.85	35.99†‡

Open Resources

- TVsub: DCU-Tencent Chinese-English Dialogue Corpus
  - https://github.com/longyuewangdcu/tvsub
- MVsub: DCU-Huawei Chinese-English Dialogue Corpus
  - https://www.computing.dcu.ie/~lwang/corpora/resource.html

Approaches for Document Level MT

- Pre-processing Approaches
- Post-processing Approaches
- RNN-based Document-Level MT Models
- Transformer-based Document-Level MT Models

Post-processing Approaches

- Elena Voita, Rico Sennrich, and Ivan Titov, *Context-Aware Monolingual Repair for Neural Machine Translation*, **EMNLP** 2019

- We propose a monolingual DocRepair model to correct inconsistencies between sentence-level translations. DocRepair performs automatic post-editing on a sequence of sentence-level translations, refining translations of sentences in context of each other.
- For training, the DocRepair model requires only monolingual document-level data in the target language.

Post-processing Approaches

Figure 1: Training procedure of DocRepair. First, round-trip translations of individual sentences are produced to form an inconsistent text fragment (in the example, both genders of the speaker and the cat became inconsistent). Then, a repair model is trained to produce an original text from the inconsistent one.

Post-processing Approaches

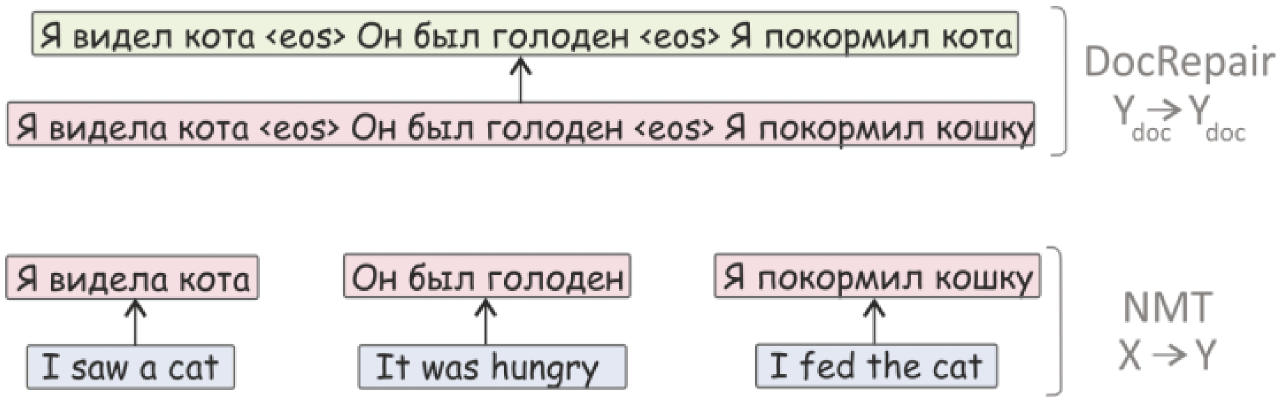


Figure 2: The process of producing document-level translations at test time is two-step: (1) sentences are translated independently using a sentence-level model, (2) DocRepair model corrects translation of the result-ing text fragment.

Post-processing Approaches

Voita et al. EMNLP 2019:

- We show that this approach successfully imitates inconsistencies we aim to fix: using contrastive evaluation, we show large improvements in the translation of several contextual phenomena in an English→Russian translation task, as well as improvements in the BLEU score.
- We also conduct a human evaluation and show a strong preference of the annotators to corrected translations over the baseline ones.
- Moreover, we analyze which discourse phenomena are hard to capture using monolingual data only.

Approaches for Document Level MT

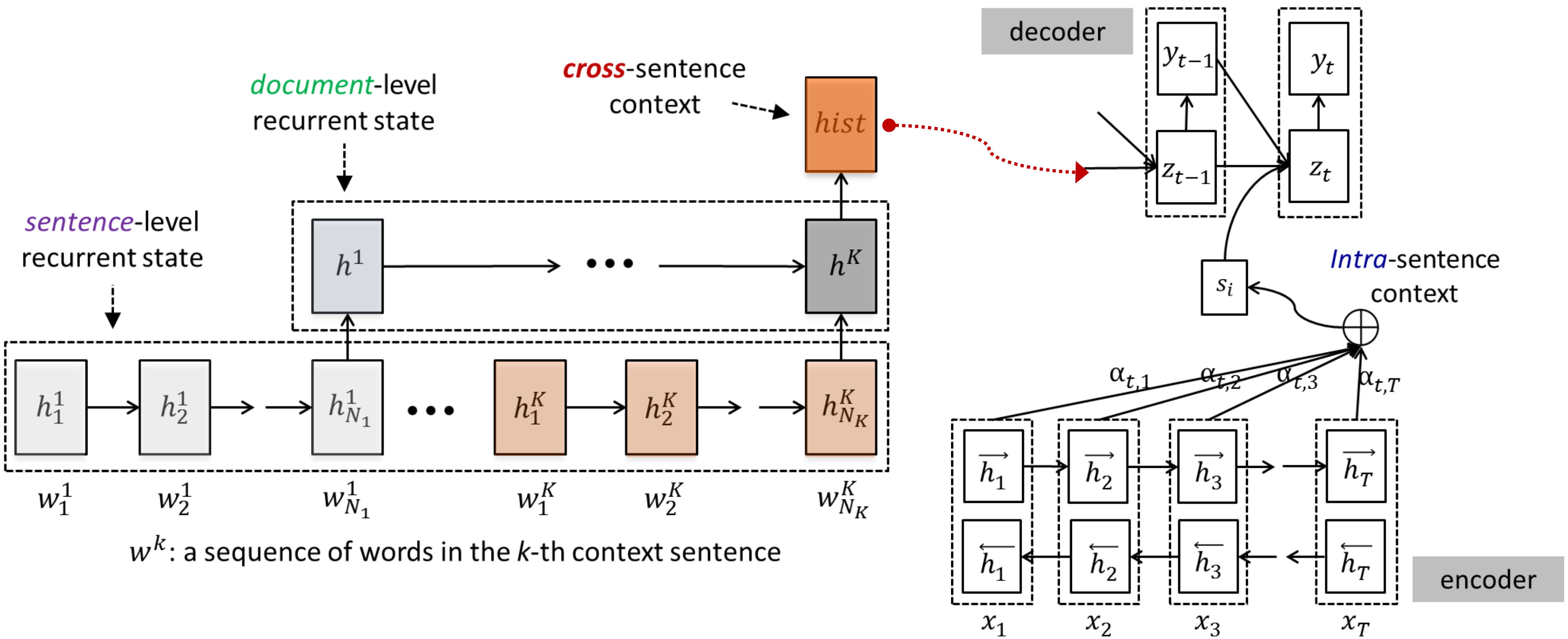
- Pre-processing Approaches
- Post-processing Approaches
- RNN-based Document-Level MT Models
- Transformer-based Document-Level MT Models

RNN-based Document Level MT Models

- Longyue Wang, Zhaopeng Tu, Andy Way, Qun Liu. *Exploiting Cross-Sentence Context for Neural Machine Translation*. **EMNLP** 2017

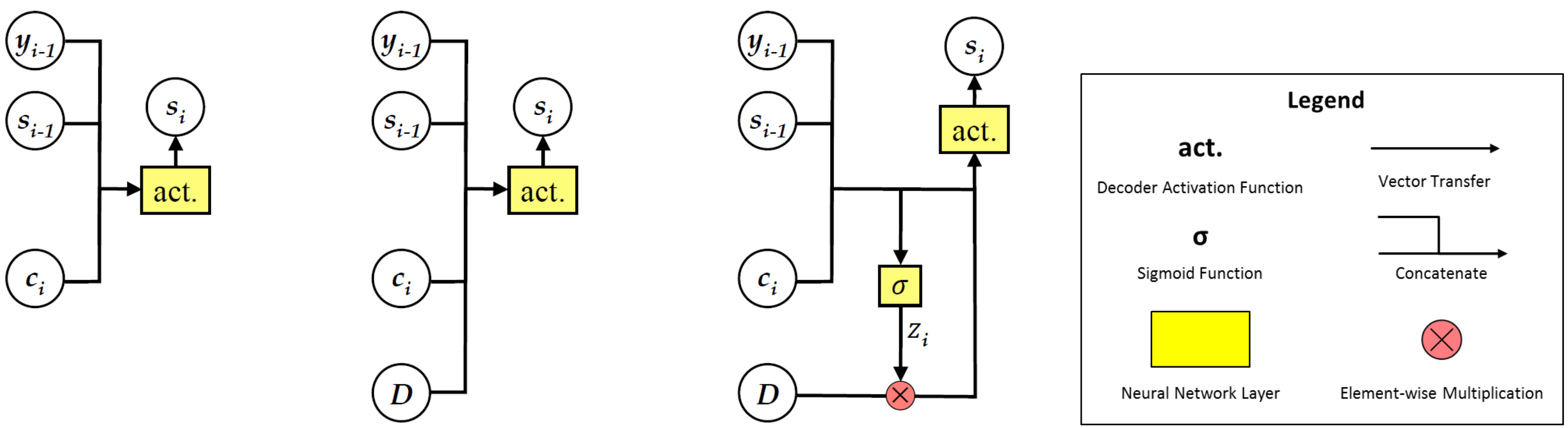
Hierarchical Recurrent Network

- Given a source sentence to be translated, we consider its K **previous sentences** in the same document as **cross-sentence context** C.
  - We first model C in a **hierarchical way**: sentence-/document-level
  - We then integrate summary of the **global context** D into NMT model



Approach

- Integrate the **historical annotations** *D* into NMT with **three strategies**:
  - **Initialization**: use *D* to initialize either encoder, decoder or both
  - **Auxiliary Context** (b): directly use *D* to work together with the dynamic intra-sentence context produced by an attention model
  - **Gating Auxiliary Context** (c): is used to dynamically control the amount of information flowing from the auxiliary context at each decoding step



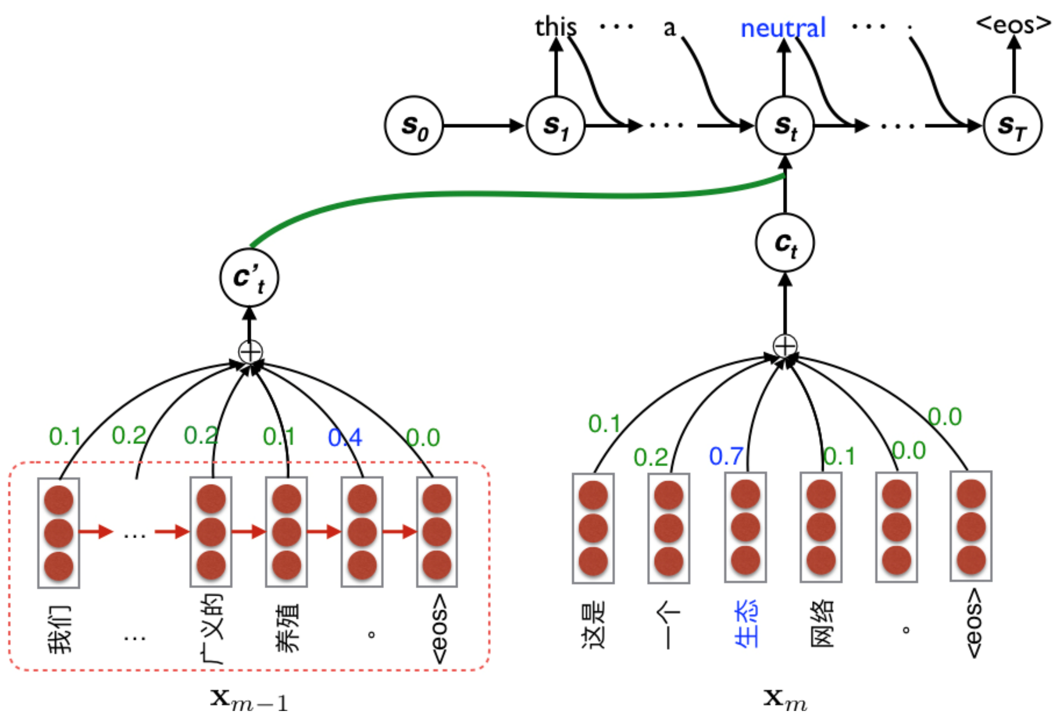
(a) standard decoder
(b) decoder with auxiliary context
(c) decoder with gating auxiliary context

RNN-based Document Level MT Models

- Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. *Does neural machine translation benefit from larger context?* **arXiv**:1704.05135

Multi-Attention

- Jean et al. (2017) propose an **additional encoder-attention** model to encode and select part of the **previous source sentence** for generating each target word.



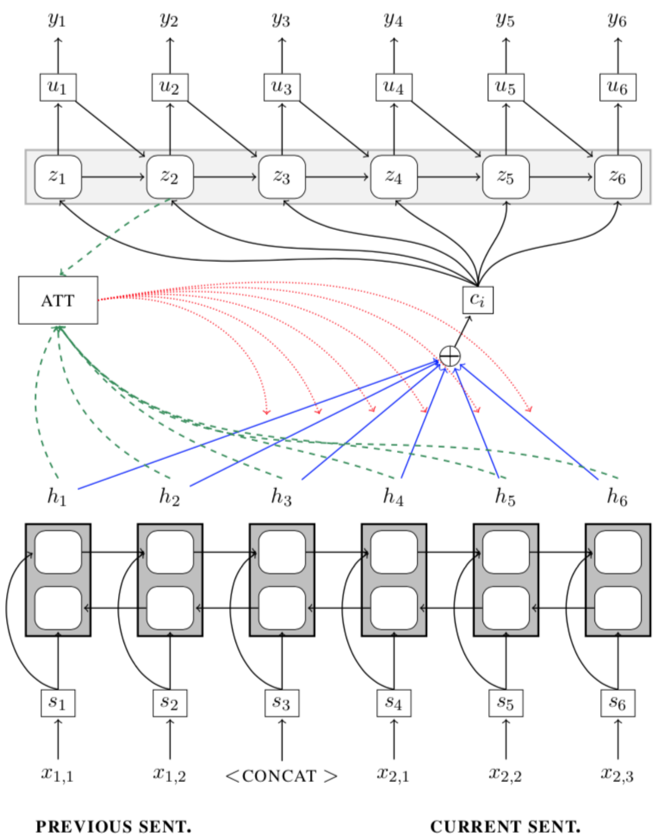
Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? arXiv:1704.05135

RNN-based Document Level MT Models

- Rachel Bawden, Rico Sennrich, Alexandra Birch, Barry Haddow, *Evaluating Discourse Phenomena in Neural Machine Translation*, **NAACL** 2018

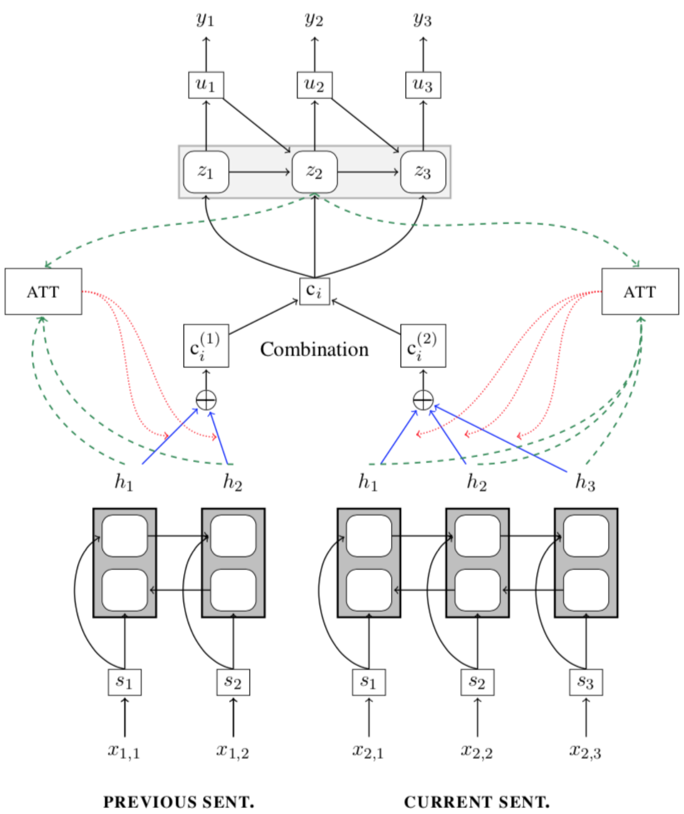
Single-Encoder Approach

- Tiedemann and Scherrer (2017) propose two context models:
  - 2-To-2
    - Trained on the concatenated source and target sentences
  - 2-To-1
    - Only concatenate source side sentences
- Both of these two models based on a single encoder model.
- Concatenate the previous sentence and current sentence with a <*CONCAT*> label.



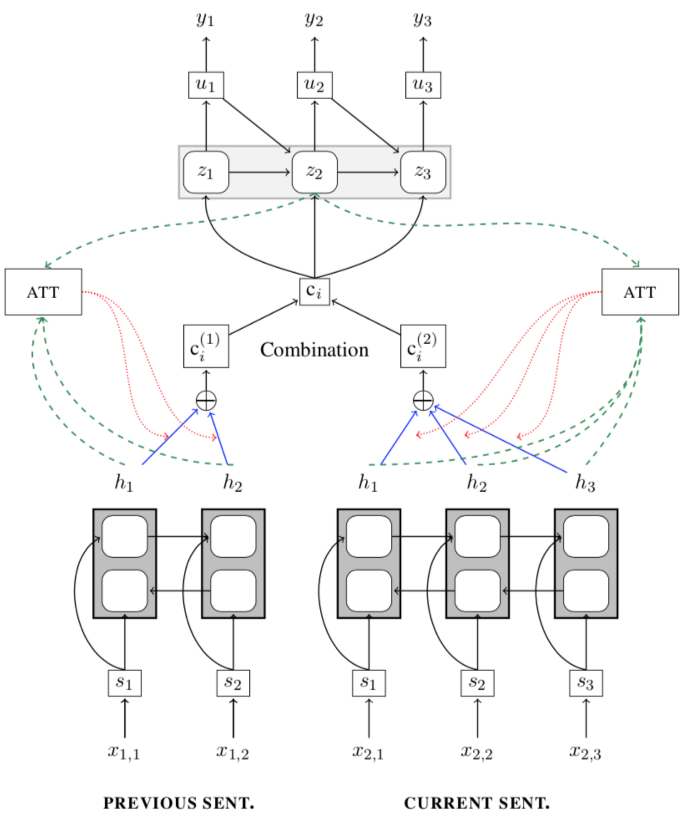
Multi-Encoder Approach

- Multi-encoder model
  – Encode the previous sentence with a separate encoder to get $c_i^{(1)}$
  – $c_i^{(2)}$ is the context vector of current sentence
  – Combine $c_i^{(1)}$ and $c_i^{(2)}$ to be used for decoding
- Three combination strategies
  – Concatenation
  – Attention gate
  – Hierarchical attention



Multi-Encoder Approach

- Three combination strategies
  - Concatenation
    - $c_i = W_c \left[c_i^{(1)}; c_i^{(2)} \right] + b$
  - Attention gate
    - $r_i = \tanh\left(W_r c_i^{(1)} + W_s c_i^{(2)} \right) + b_r$
    - $c_i = r_i \cdot W_t c_i^{(1)} + (1 - r_i) \cdot W_u c_i^{(2)})$
  - Hierarchical attention
    - $c_i = \sum_{k=1}^{K} \beta_i^{(k)} U_i^{(k)} c_i^{(k)}$
    - Where $\beta_i^{(k)}$ is the attention score

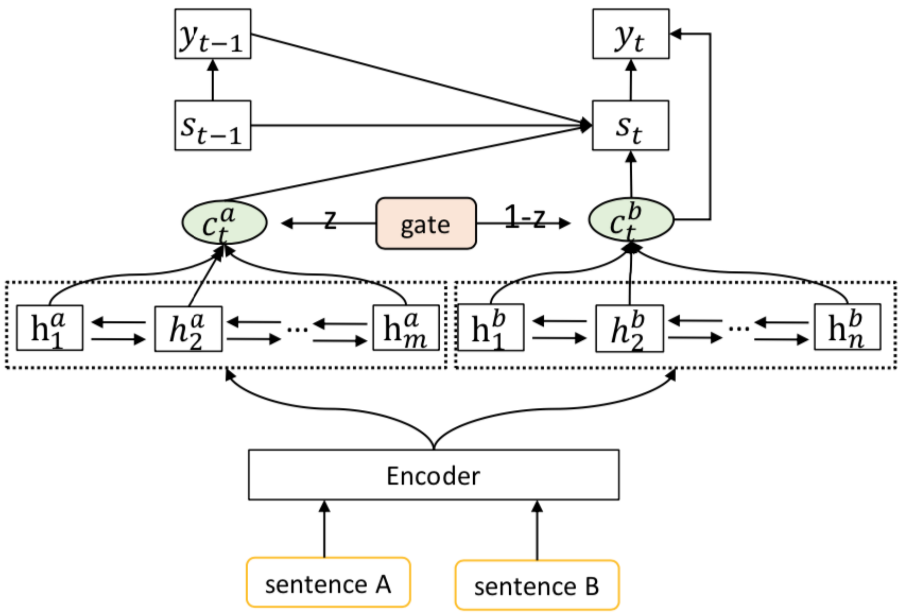


RNN-based Document Level MT Models

- Shaohui Kuang, Deyi Xiong. *Fusing Recency into Neural Machine Translation with an Inter-Sentence Gate Model*. **COLING** 2018
- Shaohui Kuang, Deyi Xiong, Weihua Luo, Guodong Zhou. *Modeling Coherence for Neural Machine Translation, with Dynamic and Topic Caches*. **COLING** 2018

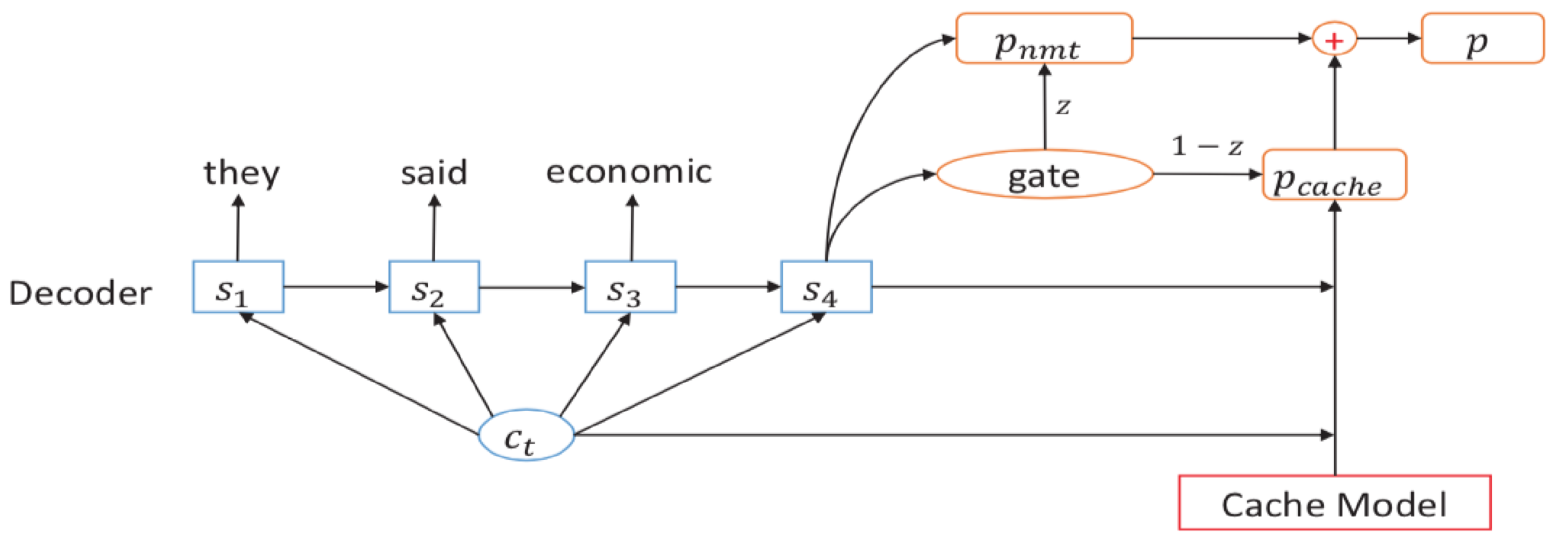
Inter-Sentence Gate Model

- Using an encoder to encode two sentences (preceding and current) at once and get the context vector c_t^a and c_t^b
- Leveraging the representation of c_t^a, c_t^b with an inter-sentence gate



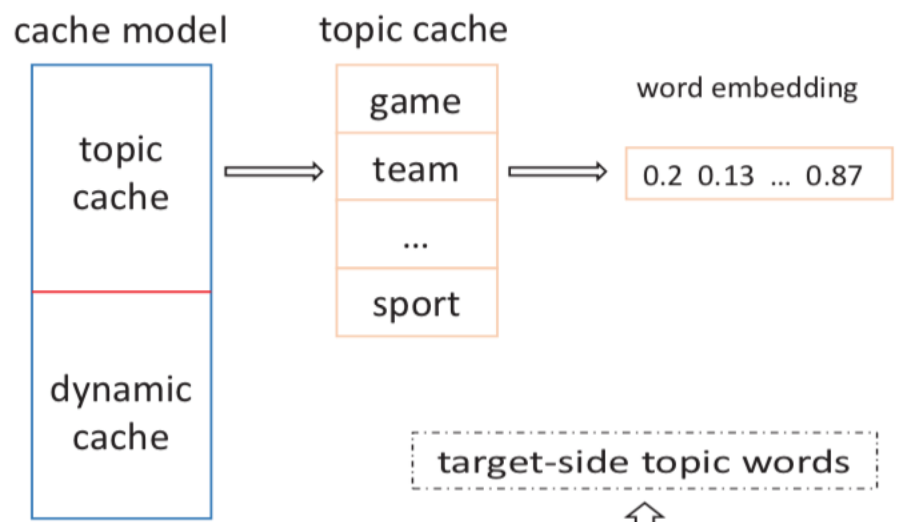
Dynamic and Topic Caches

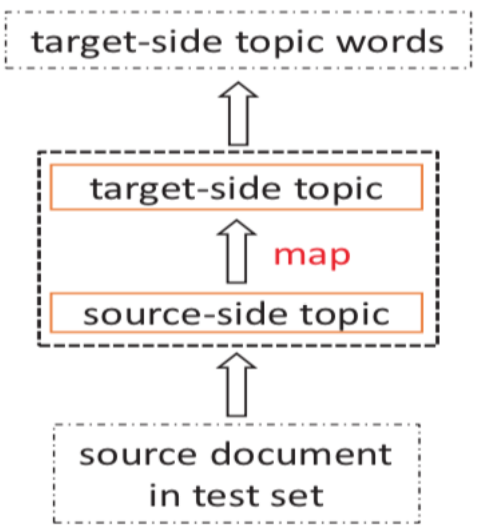
- Cache-based layer
  - At each decoding step t, we use the scorer to score y_t if it is in the cache
  - Integrating into NMT



Dynamic and Topic Caches

- Dynamic Cache
  - Extracting words from recently translated sentences and the partial translation of current sentence being translated as words of dynamic cache
- Topic Cache
  - LDA learns source- and target-side separately
  - Estimate a topic projection distribution over all target-side topics $p(z_t|z_s)$ for each source topic z_s by collecting events and accumulating counts of (z_s, z_t) from aligned document pairs.



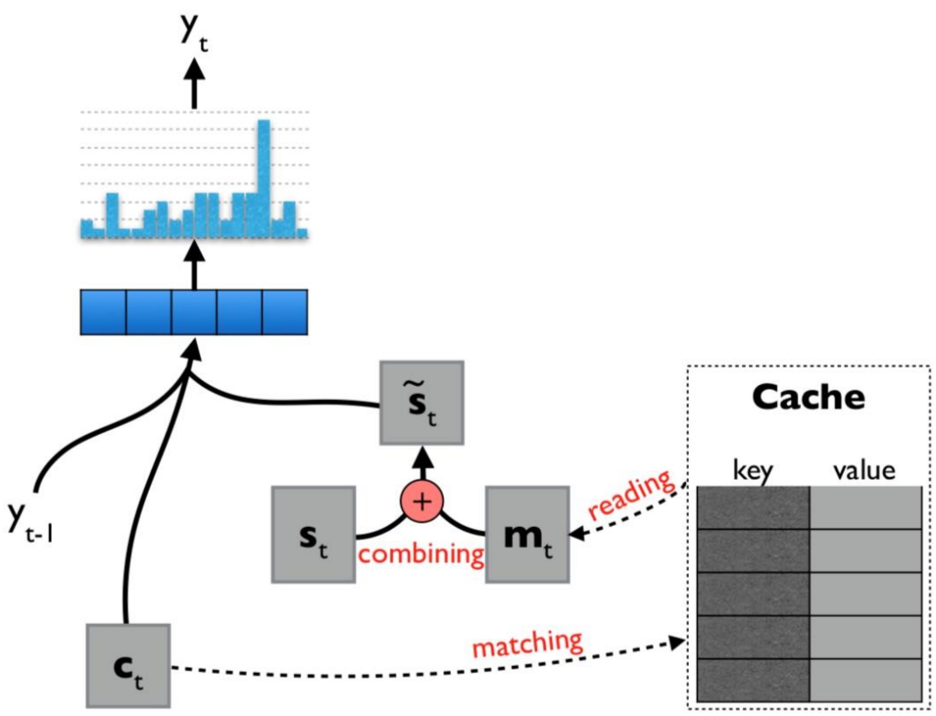


RNN-based Document Level MT Models

- Zhaopeng Tu, Yang Liu, Shuming Shi, Tong Zhang. *Learning to Remember Translation History with a Continuous Cache*. **ACL** 2018
- Zhaopeng Tu, Yang Liu, Shuming Shi, Tong Zhang. *Learning to Remember Translation History with a Continuous Cache*. **TACL** 2018

Cache Model

- Tu et al. (2018) propose to augment NMT models with a key-value memory network, which stores the translation history in terms of bilingual hidden representations at decoding steps of previous sentences.



Approaches for Document Level MT

- Pre-processing Approaches
- Post-processing Approaches
- RNN-based Document-Level MT Models
- Transformer-based Document-Level MT Models

Transformer-based Document Level MT Models

- Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, James Henderson. *Document-Level Neural Machine Translation with Hierarchical Attention Networks*. **EMNLP** 2018

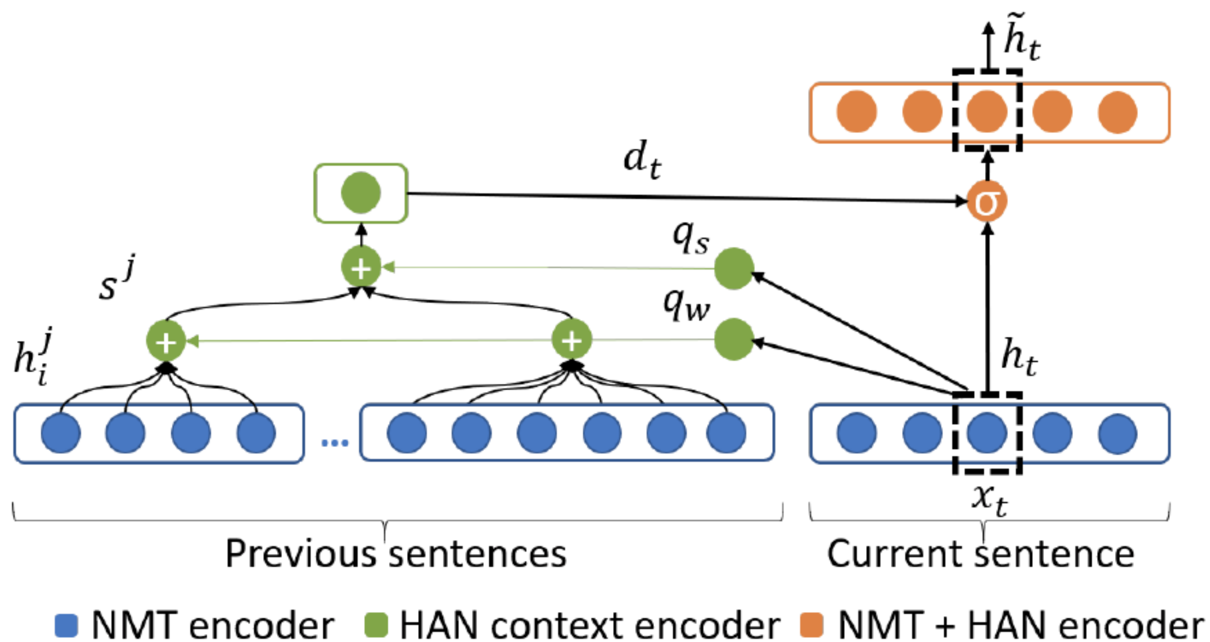
Hierarchical Attention Networks

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

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

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



Transformer-based Document Level MT Models

- Jiacheng Zhang, Huanbo Luan, Maosong Sun, FeiFei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. *Improving the Transformer Translation Model with Document-Level Context*. **EMNLP** 2018

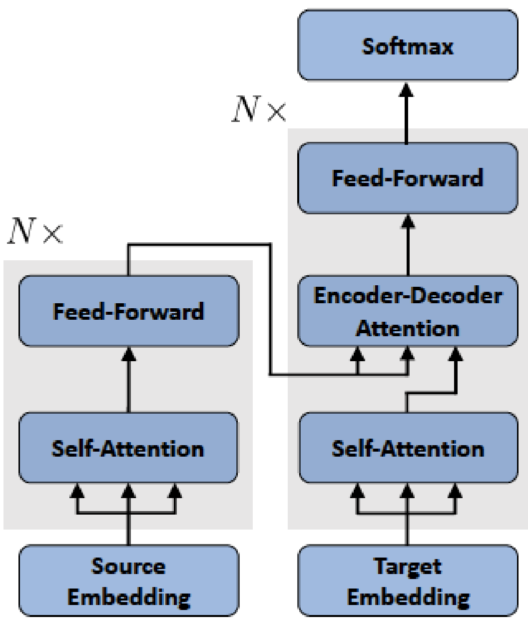
Context Encoder & Two-Step Training

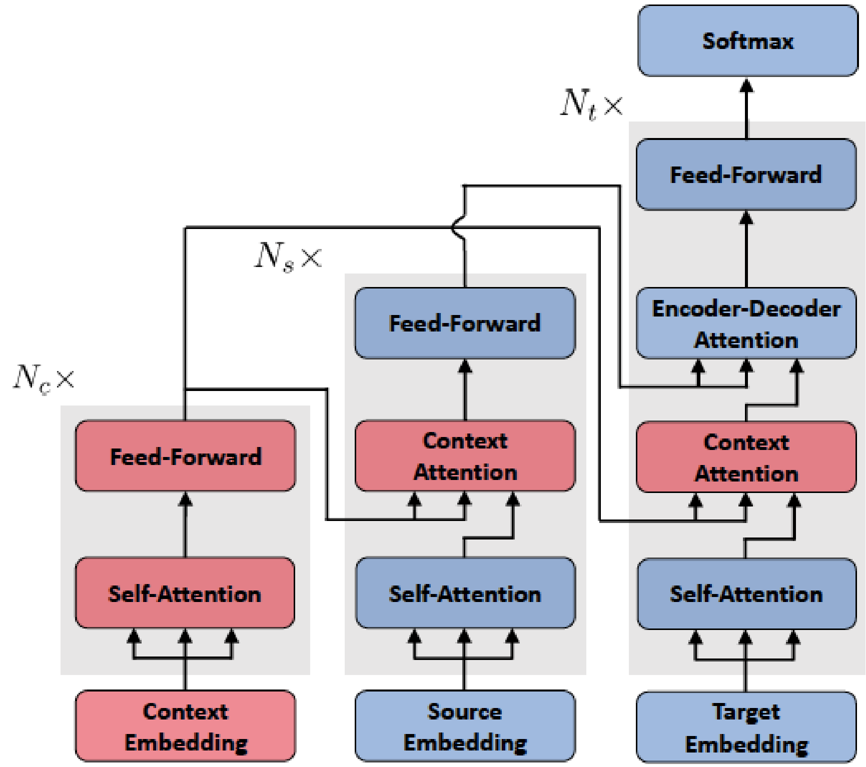
- Extend the Transformer with a new context encoder to represent document-level context.
- Introduce a two-step training method to take full advantage of abundant sentence-level parallel corpora and limited document-level parallel corpora

Context Encoder & Two-Step Training

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) = MultiHead(Xc;Xc;Xc); Q = K = V = X_c





Context Encoder & Two-Step Training

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:

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

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

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

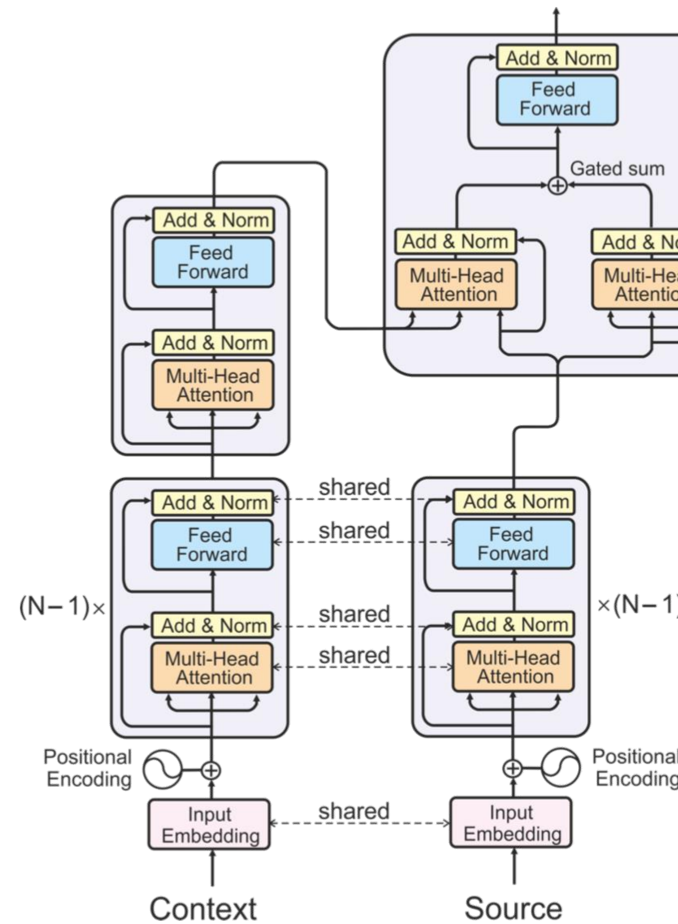
- Our approach keeps θs fixed when estimating θd

Transformer-based Document Level MT Models

- Elena Voita, Pavel Serdyukov, Rico Sennrich, Ivan Titov. *Context-Aware Neural Machine Translation Learns Anaphora Resolution*. **ACL** 2018
- Elena Voita, Rico Sennrich, Ivan Titov. *When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion*. **ACL** 2019

Context-Aware Neural Machine Translation

- First N-1 layer are shared with context sentence
- Context encoder :
  - One for encode the representation of context information
  - Other for leveraging context information $c_i^{(c)}$ and current information $c_i^{(s)}$ with a gate sum unit
- Gate sum unit:
  - $g_i = \sigma\left(\mathrm{W}_g\left[c_i^{(s)}, c_i^{(c)}\right] + b_g\right)$
  - $c_i = g_i c_i^{(s)} + (1 - g_i) c_i^{(c)}$



Transformer-based Document Level MT Models

- Liangyou Li, Qun Liu, *Pre-trained Language Models for Neural Machine Translation*, ongoing work

Background & Motivation

- Existing work usually focuses on limited context.

We use $k = 3$ previous sentences, which gave the best performance on the development set.

We consider two versions of our discourse-aware model: one using the previous sentence as the context, another one relying on the next sentence. We hypothesize that both the previous and

limited influence. Therefore, we set the number of preceding sentences to 2 in the following experiments. [5]

batch size was 80. All our models considered the previous three sentences (i.e., $K = 3$) as cross-sentence context.

- Two of the challenges when using large context
  - Performance degradation
  - Data is scarce

Method

- Idea：
  - Context manipulation
  - Pretraining + Fine-tuning
  - Multi-task learning
- Input format follows Tiedemann and Scherrer (2017)

Context:	His cat is cute
Input:	It likes fish
Extended input:	His cat is cute [SEP] It likes fish

Summary

- Using large context is under-explored in document-level NMT
- We provide our first trial to incorporate long context information without performance degrading
  - Context manipulation
  - Pre-training on monolingual data + fine-tuning
  - Multi-task learning
- Using large context is challenging
  - Higher computation cost
  - Hard to make smart selection on relevant context information
  - Parallel data is scarce
  - ……

Content

Document-Level MT Evaluation

- Elisabet Comelles, Jesús Giménez, Lluís Màrquez, Irene Castellón, Victoria Arranz. *Document-level Automatic MT Evaluation based on Discourse Representations*. **MetricsMATR** 2010
- Billy T.M. Wong, Cecilia F.K. Pun, Chunyu Kit, Jonathan J. Webster. *Lexical cohesion for evaluation of machine translation at document level*. **NLP-KE** 2011
- Billy T. M. Wong and Chunyu Kit. *Extending Machine Translation Evaluation Metrics with Lexical CohesionTo Document Level*. **EMNLP** 2012
- Rachel Bawden, Rico Sennrich, Alexandra Birch, Barry Haddow. *Evaluating Discourse Phenomena in Neural Machine Translation.* **NAACL** 2018

Document-Level MT Evaluation

- Samuel Läubli, Rico Sennrich, Martin Volk. *Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation.* **EMNLP** 2018
- Mathias Müller,Annette Rios,Elena Voita,Rico Sennrich. *A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation.* **WMT**2018
- Elena Voita, Rico Sennrich, Ivan Titov. *When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion*. **ACL** 2019

Document-Level MT Evaluation

- Marine Carpuat, Michel Simard. *The Trouble with SMT Consistency*. **WMT** 2012
- Liane Guillou. *Analysing lexical consistency in translation*. **DiscoMT** 2013
- Liane Guillou, Christian Hardmeier. *PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation*. **DiscoMT** 2016
- Pierre Isabelle, Colin Cherry, George Foster. *A Challenge Set Approach to Evaluating Machine Translation*. **EMNLP**2017

Content

1 Errors of MT at the Document Level

2 Document-Level MT Approaches

3 Document-Level MT Evaluations

4 Conclusions and Future Directions

Conclusions

- Analysis the phenomena of document level MT
- Give a taxonomy for document level MT errors
- A survey of document level MT approaches
  - Preprocessing approaches
  - Postprocessing approaches
  - RNN-based Models
  - Tranformer-based Models
  - Dealing with large context using pre-trained language models
- A list of literatures on document level MT evaluations

Suggested Future Directions

- Dealing with large context
- Dealing with very long documents
- Preprocessing and postprocessing
- Linguistically informed approaches
  - Entities and relations
  - Anaphora / Ellipse / Coreference
- Automatic Evaluation Metrics

Security Level: Public

Thank you

www.huawei.com

www.huawei.com

HUAWEI TECHNOLOGIES CO., LTD.